



The Routledge Handbook of English Language and Digital Humanities

Edited by Svenja Adolphs and Dawn Knight

The Routledge Handbook of English Language and Digital Humanities

The Routledge Handbook of English Language and Digital Humanities serves as a reference point for key developments related to the ways in which the digital turn has shaped the study of the English language and of how the resulting methodological approaches have permeated other disciplines. It draws on modern linguistics and discourse analysis for its analytical methods and applies these approaches to the exploration and theorisation of issues within the humanities.

Divided into three sections, this handbook covers:

- sources and corpora;
- analytical approaches;
- English language at the interface with other areas of research in the digital humanities.

In covering these areas, more traditional approaches and methodologies in the humanities are recast and research challenges are re-framed through the lens of the digital. The essays in this volume highlight the opportunities for new questions to be asked and long-standing questions to be reconsidered when drawing on the digital in humanities research.

This is a ground-breaking collection of essays offering incisive and essential reading for anyone with an interest in the English language and digital humanities.

Svenja Adolphs is Professor of English Language and Linguistics at the University of Nottingham, UK. Her research interests are in the areas of corpus linguistics (in particular, multimodal spoken corpus linguistics), pragmatics and discourse analysis. She has published widely in these areas, including *Introducing Electronic Text Analysis* (2006, Routledge), *Corpus and Context: Investigating Pragmatics Functions in Spoken Discourse* (2008), *Introducing Pragmatics in Use* (1st ed. 2011, 2nd ed. 2020, Routledge, with Anne O’Keeffe and Brian Clancy) and *Spoken Corpus Linguistics: From Monomodal to Multimodal* (2013, Routledge, with Ronald Carter).

Dawn Knight is Reader in Applied Linguistics at Cardiff University. Her research interests lie in the areas of corpus linguistics, discourse analysis, digital interaction, non-verbal communication and the sociolinguistic contexts of communication. The main contribution of her work has been to pioneer the development of a new research area in applied linguistics: multimodal corpus-based discourse analysis. Dawn is the principal investigator on the ESRC/AHRC-funded CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh) project (2016–2020) and is currently the chair of the British Association for Applied Linguistics (BAAL), representing over one thousand applied linguists within the UK (2018–2021).

Routledge Handbooks in English Language Studies

Routledge Handbooks in English Language Studies provide comprehensive surveys of the key topics in English language studies. Each Handbook focuses in detail on one area, explaining why the issue is important and then critically discussing the leading views in the field. All the chapters are specially commissioned and written by prominent scholars. Coherent, accessible and carefully compiled, *Routledge Handbooks in English Language Studies* are the perfect resource for both advanced undergraduate and postgraduate students.

THE ROUTLEDGE HANDBOOK OF STYLISTICS

Edited by Michael Burke

THE ROUTLEDGE HANDBOOK OF LANGUAGE AND CREATIVITY

Edited by Rodney H. Jones

THE ROUTLEDGE HANDBOOK OF CONTEMPORARY ENGLISH PRONUNCIATION

Edited by Okim Kang, Ron Thomson and John M. Murphy

THE ROUTLEDGE HANDBOOK OF ENGLISH LANGUAGE STUDIES

Edited by Philip Seargeant, Ann Hewings and Stephen Pihlaja

THE ROUTLEDGE HANDBOOK OF ENGLISH LANGUAGE AND DIGITAL HUMANITIES

Edited by Svenja Adolphs and Dawn Knight

The Routledge Handbook of English Language and Digital Humanities

Edited by Svenja Adolphs and Dawn Knight

First published 2020
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
52 Vanderbilt Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2020 selection and editorial matter, Svenja Adolphs & Dawn Knight; individual chapters, the contributors

The right of Svenja Adolphs & Dawn Knight to be identified as the authors of the editorial material, and of the authors for their individual chapters, has been asserted in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data

Names: Adolphs, Svenja, editor. | Knight, Dawn, 1982– editor.

Title: The Routledge handbook of English language and digital humanities / edited by
Svenja Adolphs and Dawn Knight.

Description: Abingdon, Oxon ; New York, NY : Routledge, 2020. | Includes
bibliographical references and index.

Identifiers: LCCN 2019051030 | ISBN 9781138901766 (hardback)

Subjects: LCSH: English language—Research—Data processing. | English language—
Research—Methodology. | Linguistics—Research—Data processing. | Linguistics—
Research—Methodology. | Digital humanities.

Classification: LCC PE1074.5 .R68 2020 | DDC 420.72—dc23

LC record available at <https://lcn.loc.gov/2019051030>

ISBN: 978-1-138-90176-6 (hbk)

ISBN: 978-1-003-03175-8 (ebk)

Typeset in Times New Roman
by Apex CoVantage, LLC

Contents

<i>List of figures</i>	<i>viii</i>
<i>List of tables</i>	<i>xi</i>
<i>List of contributors</i>	<i>xiii</i>
<i>Acknowledgements</i>	<i>xxi</i>
1 English language and the digital humanities <i>Svenja Adolphs and Dawn Knight</i>	1
2 Spoken corpora <i>Karin Aijmer</i>	5
3 Written corpora <i>Sheena Gardner and Emma Moreton</i>	26
4 Digital interaction <i>Jai Mackenzie</i>	49
5 Multimodality I: speech, prosody and gestures <i>Phoebe Lin and Yaoyao Chen</i>	66
6 Multimodality II: text and image <i>Sofia Malamatidou</i>	85
7 Digital pragmatics of English <i>Irma Taavitsainen and Andreas H. Jucker</i>	107
8 Metaphor <i>Wendy Anderson and Elena Semino</i>	125
9 Grammar <i>Anne O’Keeffe and Geraldine Mark</i>	143

Contents

10	Lexis	164
	<i>Marc Alexander and Fraser Dallachy</i>	
11	Ethnography	185
	<i>Piia Varis</i>	
12	Mediated discourse analysis	202
	<i>Rodney H. Jones</i>	
13	Critical discourse analysis	220
	<i>Paul Baker and Mark McGlashan</i>	
14	Conversation analysis	242
	<i>Eva Maria Martika and Jack Sidnell</i>	
15	Cross-cultural communication	263
	<i>Eric Friginal and Cassie Dorothy Leymarie</i>	
16	Sociolinguistics	283
	<i>Lars Hinrichs and Axel Bohmann</i>	
17	Literary stylistics	306
	<i>Michaela Mahlberg and Viola Wiegand</i>	
18	Historical linguistics	328
	<i>Freek Van de Velde and Peter Petré</i>	
19	Forensic linguistics	360
	<i>Nicci MacLeod and David Wright</i>	
20	Corpus linguistics	378
	<i>Gavin Brookes and Tony McEnery</i>	
21	English language and classics	405
	<i>Alexandra Trachsel</i>	
22	English language and history: geographical representations of poverty in historical newspapers	418
	<i>Ian N. Gregory and Laura L. Paterson</i>	
23	English language and philosophy	440
	<i>Jonathan Tallant and James Andow</i>	

24	English language and multimodal narrative <i>Riki Thompson</i>	456
25	English language and digital literacies <i>Paul Spence</i>	472
26	English language and English literature: new ways of understanding literary language using psycholinguistics <i>Kathy Conklin and Josephine Guy</i>	494
27	English language and digital health humanities <i>Brian Brown</i>	511
28	English language and public humanities: ‘An army of willing volunteers’: analysing the language of online citizen engagement in the humanities <i>Ben Clarke, Glenn Hadikin, Mario Saraceni and John Williams</i>	530
29	English language and digital cultural heritage <i>Lorna M. Hughes, Agiatis Benardou and Ann Gow</i>	555
30	English language and social media <i>Caroline Tagg</i>	568
	<i>Index</i>	587

Figures

3.1	Frequency and MI of top ten collocates of ‘humanity’ in COCA	37
3.2	Raw frequency of ‘humanities’ in COHA	37
3.3	Collocates of ‘humanities’ in GloWbE	38
3.4	TEI encoding for capturing information about the sender, recipient, origin, destination and date, using metadata from the MCC as an example	40
3.5	Sample concordance lines for ‘I hope you’	44
4.1	Mumsnet logo	54
4.2	Research design for the Mumsnet study	55
5.1	A concordance of ‘I think’ generated using Lin’s (2016) web-as-multimodal-corpus concordancing tool	76
5.2	A screenshot of ELAN (version 4.9.4) showing tiers of information generated using an automatic tool	77
6.1	Parameter and values	97
Image 6.1	Screenshot of Visual Tagger tagging interface	98
Image 6.2	Screenshot of Visual Tagger search interface	98
7.1	Three circles of corpus-based speech act retrieval according to Jucker and Taavitsainen 2013: 95	115
8.1	Metaphor Map of English showing connections between category 1E08 Reptiles and all of the macro-categories	136
10.1	Section of the Historical Thesaurus category ‘03.12.15 Money’ from the Thesaurus’ website	177
10.2	Metaphor ‘card’ showing links between ‘Money’ and ‘Food and eating’ from the Mapping Metaphor project website	179
10.3	Top search results for semantic category ‘BJ : 01 : m – Money’ in the semantically tagged Hansard Corpus	181
13.1	Process for extracting articles and comments corpora	225
15.1	Comparison of average agent dimension scores for ‘addressee-focused, polite and elaborated speech’ vs. ‘involved and simplified narrative’	272
16.1	Sociolinguistics (here represented as ‘LVC’, i.e. ‘language variation and change studies’) versus other branches of empirical linguistics	287
16.2	Factor 5 score distribution for different genres	297
Concordance 17.1	All 32 instances of <i>fog</i> in <i>Bleak House</i> , retrieved with CLiC	315
17.1	<i>Fog</i> collocates in <i>Bleak House</i> retrieved with the ‘links’ function of Voyant Tools (top) and their context illustrated with the CLiC KWICGrouper (bottom), both within a context of five words around <i>fog</i>	316

17.2	WordWanderer collocate visualisation of <i>fog</i> in <i>Bleak House</i> overall (top panel) and in the comparison view with <i>lord</i> (bottom panel)	317
Concordance 17.2	Sample of instances of <i>once upon a time</i> in the 19th Century Children's Literature Corpus (ChiLit; 10 out of 21 examples)	319
Concordance 17.3	All instances of <i>put the case that</i> in the CLiC corpora (all appear in <i>Great Expectations</i>)	320
Concordance 17.4	Random sample of <i>it seems to me that</i> in the BNC Spoken (from a total of 161 hits in 77 different texts)	321
Concordance 17.5	Sample lines of <i>it seems to me that</i> in the 19C corpus (10 out of 32 examples)	322
18.1	Frequency history of 'weorðan' in two different corpora	335
18.2	Genitive alternation in the BROWN and F-BROWN corpora	339
18.3	Cohen-friendly association plot of the genitive alternation in British and American English in the 1960s and the 1990s	339
18.4	Distribution of 'weorðan' and 'wesan' as passive auxiliaries in Old English	341
18.5	Increasing use of adjectival modifiers with intensifying 'such'	342
18.6	AP vs. NP in secondary predicates	343
18.7	The genitive alternance	345
18.8	A mixed model for the diachrony genitive alternation	346
18.9	Diachrony of the conjunction <i>albeit</i> in the HANSARD corpus	347
18.10	Correlograms of the diachrony of conjunction <i>albeit</i> in the HANSARD corpus	348
18.11	Detrending the time series of conjunction <i>albeit</i> in the HANSARD corpus by working in differences	349
18.12	Forecasting the diachrony of <i>albeit</i> on the basis of the ARIMA model	350
18.13	Increased grammaticalisation of 'be going to' in Early Modern English	351
18.14	VNC analysis of increased grammaticalisation of 'be going to' in Early Modern English	352
20.1	Frequency of gendered nouns in Michel et al. (2011) and Baker (2014)	387
20.2	First-order collocates of 'rude'	396
20.3	Second-order collocates when the search word is 'receptionists'	397
20.4	Concordance lines for collocation of 'rude' with 'receptionists'	398
22.1	Example PNC from <i>The Times</i> 1940s corpus	423
22.2	Example concordance lines for <*poverty*> in <i>The Times</i> 1990s corpus	424
22.3	Dot maps showing the pattern of PNCs in the 1940s (left) and 1980s (right)	426
22.4	Density smoothed map of all <*poverty*> PNCs from <i>The Times</i> , 1940–2009	428
22.5	Macro-geographies of <*poverty*> instances over time, 1950s onward	429
22.6	Spatial segregation analysis comparing the locations of PNCs in the 1980s with those from other decades from the 1950s on. PNCs and densities are for the 1980s	434
22.7	Spatial segregation analysis comparing the locations of PNCs from the 1980s onwards with those from the 1950s to 1970s. PNCs and densities are for the 1980s onwards	435
24.1	<i>Have You Met Olivia?</i>	462
24.2	RSAnimate – <i>Changing the Education Paradigm</i> (2010a)	465

Figures

24.3	RAnimate – <i>Economics Is for Everyone!</i> (2016)	465
24.4	RAnimates: <i>Smile or Die</i> (2010b); <i>The Paradox of Choice</i> (2011); <i>How to Help Every Child Reach Their Potential</i> (2015)	467
28.1	The traditional model for the production and diffusion of knowledge	531
28.2	The ‘citizen’ model for the production and diffusion of knowledge	533
28.3	WordSmith tools pattern view for the node ‘I’ in the Ancient Lives corpus	547
28.4	Sample concordance of ‘I am’ in the Ancient Lives corpus. (The data have been selected to include the set of seven occurrences of ‘I am not’.)	548
28.5	WordSmith Tools pattern view for the node ‘we’ in the Ancient Lives corpus	549
28.6	Full concordance of ‘we are’ in the Ancient Lives corpus	550
30.1	Posted in group labelled ‘Eating and drinking’, four participants	577
30.2	Individual interaction with S	578
30.3	Emoji posted in group ‘Eating and drinking’, four participants	580

Tables

3.1	Types of corpus exemplified	28
3.2	Concordance lines of ‘humanity’ in the discipline of classics in the BAWE corpus	36
3.3	‘So-called humanities’ in COCA	38
3.4	Mean scores for dimensions across five datasets	41
3.5	Raw and normalised frequencies (per million) for Pronouns, I + Verb, and I + Verb + You across the MCC, BTCC and LALP corpus	42
3.6	Instances of I + Verb + You across the three corpora, organised by frequency	43
6.1	Multimodal concordance	91
6.2	Hypothetical values on four variables	94
6.3	Distribution of animate and inanimate participants in <i>Scientific American</i> and <i>La Recherche</i>	99
6.4	Distribution of specific animate and inanimate participants in <i>Scientific American</i> and <i>La Recherche</i>	99
6.5	Distribution of verbs and nouns in <i>Scientific American</i> and <i>La Recherche</i>	100
8.1	Metaphorical connections with 1E08 Reptiles as target	137
8.2	Sample metaphorical connections with 1E08 Reptiles as source	138
9.1	Breakdown of text types in the London-Lund Corpus	147
9.2	Frequencies (per million words) of the modal verb <i>must</i> in affirmative and negative patterns in a sample of the CLC across the six levels of the CEFR	155
9.3	A summary of the development of competencies in form (shaded) and function of affirmative and negative patterns of <i>must</i> , across forms and functions, taken from the English Grammar Profile	156
13.1	Keywords in articles and comments	227
13.2	Concordance table of ten citations of the keyword <i>influx</i> in the Articles corpus	227
13.3	Collocates of <i>Romanians</i>	230
13.4	Verb collocates of <i>us</i> and <i>them</i> in the comments corpus	233
13.5	The most prototypical articles in the corpus	235
15.1	Linguistic dimensions of outsourced call centre discourse	271
16.1	The combined ICE and Twitter corpus in Bohmann (2019)	295
20.1	Top ten words in the GP comments, ranked by frequency	391
20.2	Top ten keywords in the GP comments, ranked by log ratio	393

Tables

20.3	Collocates of ‘rude’, ranked by MI	395
22.1	Hits for <*poverty*> in <i>The Times</i> corpora	421
22.2	Instances of <*poverty*> in <i>The Times</i> corpus by decade (pmw)	425
22.3	PNC keywords derived from a comparison of the PNCs for Greater London with those for elsewhere from the 1950s onwards	431
27.1	Keywords by categorisation of health themes in adolescent health emails	521
27.2	Examples of messages containing the common collocates ‘help’ and ‘stop’	522

Contributors

Svenja Adolphs is Professor of English Language and Linguistics at the University of Nottingham, UK. Her research interests are in the areas of corpus linguistics (in particular, multimodal spoken corpus linguistics), pragmatics and discourse analysis. She has published widely in these areas, including *Introducing Electronic Text Analysis* (2006, Routledge), *Corpus and Context: Investigating Pragmatics Functions in Spoken Discourse* (2008, John Benjamins), *Introducing Pragmatics in Use* (1st ed. 2011, 2nd ed. 2020, Routledge, with Anne O’Keeffe and Brian Clancy) and *Spoken Corpus Linguistics: From Monomodal to Multimodal* (2013, Routledge, with Ronald Carter).

Karin Aijmer is Professor Emerita in English Linguistics at the University of Gothenburg, Sweden. Her research interests focus on pragmatics, discourse analysis, modality, corpus linguistics and contrastive analysis. Her books include *Conversational Routines in English: Convention and Creativity* (1996), *English Discourse Particles: Evidence from a Corpus* (2002) and *The Semantic Field of Modal Certainty: A Study of Adverbs in English* (2007, with Anne-Marie Simon-Vandenberg).

Marc Alexander is Professor of English Linguistics at the University of Glasgow and is the third director of the *Historical Thesaurus of English*. He works primarily on the study of words, meaning and effect in English, often using data from the *Thesaurus* or approaches from the digital humanities, and his other research interests centre around meaning, rhetoric and reader manipulation in written texts.

Wendy Anderson is Professor of Linguistics at the University of Glasgow. She was Principal Investigator of the ‘Mapping Metaphor with the Historical Thesaurus’ and ‘Metaphor in the Curriculum’ projects and is associate editor of *Digital Scholarship in the Humanities* (Oxford Journals). Her research interests also include corpus linguistics and intercultural language education.

James Andow is Lecturer in Philosophy at the University of East Anglia. He previously taught at the University of Reading and the University of Nottingham. His research focuses on issues concerning the methods and epistemology of philosophy. He has published empirical work in philosophy, drawing on methods from corpus linguistics and social psychology.

Paul Baker is Professor of English Language at the Department of Linguistics and English Language, Lancaster University, where he is a member of the ESRC Centre for Corpus

Approaches to Social Science (CASS). He has written 16 books, including *Using Corpora to Analyse Discourse* (2006), *Sexed Texts: Language, Gender and Sexuality* (2008) and *Discourse Analysis and Media Attitudes* (2013). He is also commissioning editor of the journal *Corpora* (EUP).

Agiatis Benardou is Senior Research Associate in the Digital Curation Unit, ATHENA Research Center, and a research fellow at the Department of Informatics, Athens University of Economics and Business. An archaeologist by training, she has carried out extensive research and published in the field of digital methods and user needs, and has been involved in various EU and international initiatives in the digital humanities. She has taught digital curation at Panteion University, Athens, and is teaching digital methods at the Athens University of Economics and Business.

Axel Bohmann is Assistant Professor of English Linguistics at Albert-Ludwigs-Universität Freiburg. In his PhD dissertation (2017, Austin, Texas) he performed a feature-aggregation-based corpus study which develops ten fundamental dimensions of linguistic variation in English worldwide. His research interests include language variation, world Englishes, corpus linguistics and the sociolinguistics of globalisation.

Gavin Brookes is Senior Research Associate in the ESRC Centre for Corpus Approaches to Social Science (CASS) at Lancaster University. He has published research in the areas of corpus linguistics, (critical) discourse studies, multimodality and health(care) communication. He is associate editor of the *International Journal of Corpus Linguistics*, published by John Benjamins.

Brian Brown is Professor of Health Communication in the Faculty of Health and Life Sciences, De Montfort University. His interests embrace social theory as applied to health studies; the experience of health and illness from practitioner, client and carer perspectives; and new directions in the study of language in healthcare.

Yaoyao Chen is Assistant Professor in Applied Linguistics at the Macau University of Science and Technology. Her research explores multimodal corpus-based approaches for investigating the temporal and functional relationship between speech and gesture. Her research interests include multimodal corpus linguistics, language and gesture, multimodal communication and multimodal corpus pragmatics.

Ben Clarke is Senior Lecturer in Communication Studies at the University of Gothenburg. His research interests include pragmatics, functional grammar, multimodality, media and political communication and short-term diachronic change. He has published papers on these topics and, methodologically, views moving back and forth between discourse analysis and corpus linguistics as advantageous.

Kathy Conklin is Professor of Psycholinguistics at the University of Nottingham. She uses a number of behavioural and neurophysiological techniques in research examining the processing of language in well-controlled experimental contexts and, more recently (and in collaboration with Jo Guy), using authentic texts to address questions about literariness. She is co-author of the Cambridge University Press *An Introduction to Eye-Tracking: A Guide for Applied Linguistics Research*.

Fraser Dallachy is Lecturer in the *Historical Thesaurus of English* at the University of Glasgow. He has worked on semantic annotation and historical semantics projects using *Historical Thesaurus* data and is deputy director of the *Thesaurus*.

Eric Friginal is Professor of Applied Linguistics at the Department of Applied Linguistics and English as a Second Language at Georgia State University (GSU) and Director of International Programmes at the College of Arts and Sciences of GSU. He specialises in applied corpus linguistics, discourse analysis, sociolinguistics, cross-cultural communication and the analysis of spoken professional and NNS discourse.

Sheena Gardner is Professor of Applied Linguistics in the School of Humanities and the Research Centre for Arts, Memory and Communities at Coventry University. She draws on systemic-functional linguistics and multidimensional analysis in her research on genres, registers and lexico-grammar across disciplinary contexts in the BAWE corpus of university student writing.

Ann Gow is Senior Lecturer in Information Studies at the University of Glasgow, where she leads the MA (Hons) programme Digital Media and Information Studies. She specialises in the pedagogy of teaching digital media and digital curation knowledge exchange with the commercial and heritage sector. Gow leads the Learning and Teaching Strategy in the School of Humanities at UofG. She has been involved in the leadership of UK, EU and international digital heritage and digital curation projects.

Ian N. Gregory is Professor of Digital Humanities at Lancaster University. His research centres on the use of digital geographical approaches across the humanities and social sciences covering fields ranging from modern poverty, to nineteenth-century public health, to literary history.

Josephine Guy is Professor of Modern Literature at the University of Nottingham. She has published widely on nineteenth-century literary and intellectual history, on the theory and practice of text editing, on literary theory and, more recently (and in collaboration with Kathy Conklin), on the application of psycholinguistics to questions about literariness. She is currently editing two of Oscar Wilde's plays as part of the Oxford English Texts Edition of *The Complete Works of Oscar Wilde*.

Glenn Hadikin is Senior Lecturer in English Language and Linguistics at the University of Portsmouth. He has published in a wide range of journals such as *English Today*, *Corpora* and the *British Journal of Management* and currently works on an Interreg Europe project that supports women in the tech industries.

Lars Hinrichs is Associate Professor of English Language and Linguistics at the University of Texas at Austin. His doctoral work (PhD 2006, Freiburg, Germany) was a discourse analysis of Creole-English code-switching in Jamaican emails. In other work, he studies variation in speech among the Jamaican-Canadian community, variation in Standard English corpora and present-day Texas English.

Lorna M. Hughes is Professor of Digital Humanities at the University of Glasgow. Her research addresses the creation and use of digital cultural heritage for research, with a focus

on collaborations between the humanities and scientific disciplines. Lorna has worked in digital humanities and digital collections at a number of organisations in the United States and UK, including NYU, Oxford University, King's College London, the University of London School of Advanced Study and the University of Wales/National Library of Wales. She has had leadership roles in over 20 funded research projects, including *The Snows of Yesteryear: Narrating Extreme Weather* (eira.llgc.org.uk); the digital archive *The Welsh Experience of the First World War* (cymruww1.llgc.org.uk); the EPSRC-AHRC *Scottish National Heritage Partnership*; the *Living Legacies 1914–18 Engagement Centre*, *Listening and British Cultures: Listeners' Responses to Music in Britain, c. 1700–2018*; and the EU *DESIR* project (DARIAH Digital Sustainability).

Rodney H. Jones is Professor of Sociolinguistics and head of the Department of English Language and Applied Linguistics at the University of Reading. His research interests include health communication and the ways technologies of all kinds affect human interaction. His recent books include *Health Communication: An Applied Linguistic Perspective* (Routledge, 2013), *Discourse and Digital Practices* (with Christoph Hafner and Alice Chik, Routledge, 2015) and *Spoken Discourse* (Bloomsbury, 2016).

Andreas H. Jucker is Professor of English Linguistics at the University of Zurich. His current research focuses on historical pragmatics, and in particular, on the history of specific speech acts, such as greetings or apologies, and on the development of politeness and impoliteness in English.

Dawn Knight is Reader in Applied Linguistics at Cardiff University. Her current research interests lie predominantly in the areas of corpus linguistics, discourse analysis, digital interaction, non-verbal communication and the sociolinguistic contexts of communication. The main contribution of her work has been to pioneer the development of a new research area in applied linguistics: multimodal corpus-based discourse analysis. This has included the introduction of a novel methodological approach to the analysis of the relationships between language and gesture-in-use based on large-scale, real-life records of interactions (corpora). Dawn was the principal investigator on the ESRC/AHRC-funded CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh) project.

Cassie Dorothy Leymarie is the curriculum and assessment coordinator at the Global Village Project, a school for refugee girls in Decatur, Georgia. She received her PhD from the Department of Applied Linguistics and English as a Second Language at Georgia State University. She specialises in sociolinguistic inquiry and the study of the linguistic landscape.

Phoebe Lin is an Assistant Professor at The Hong Kong Polytechnic University. Her research investigates aspects of formulaic language prosody using a range of methods, including controlled experiments, corpus analysis, statistical modelling and software development. Her recent work involves developing intelligent, web-based applications that exploit the potential of YouTube as a resource for multimodal corpus research and for second language vocabulary acquisition.

Jai Mackenzie is a Research Fellow in the School of English at the University of Nottingham. Her primary research interests lie in explorations of language, gender, sexuality and parenthood, especially in digital contexts. She recently published her first monograph: *Language, Gender and Parenthood Online*.

Nicci MacLeod is Senior Lecturer in English Language and Linguistics in the Department of Humanities at Northumbria University and a freelance forensic linguistic consultant. She holds a PhD in forensic linguistics from Aston University and has published widely on the topics of police interviewing and the language of online undercover operations against child sex abusers. She is co-author with Professor Tim Grant of *Language and Online Identities: The Undercover Policing of Sexual Crime*, due in February 2020 from Cambridge University Press.

Michaela Mahlberg is Professor of Corpus Linguistics at the University of Birmingham, UK, where she is also the director of the Centre for Corpus Research. She is the editor of the *International Journal of Corpus Linguistics* (John Benjamins) and together with Wolfgang Teubert she edits the book series *Corpus and Discourse* (Bloomsbury). One of her main areas of research is Dickens's fiction and the sociocultural context of the nineteenth century. Her publications include *Corpus Stylistics and Dickens's Fiction* (Routledge, 2013), *English General Nouns: A Corpus Theoretical Approach* (John Benjamins, 2005) and *Text, Discourse and Corpora: Theory and Analysis* (Continuum, 2007, co-authored with Michael Hoey, Michael Stubbs and Wolfgang Teubert).

Sofia Malamatidou is Lecturer in Translation Studies at the University of Birmingham, UK. She has published on the topic of translation-induced linguistic change in the target language, with reference to the genre of popular science, and on a new methodological framework called corpus triangulation. Her recent research interests also include the translation of the language of promotion with regard to tourism texts.

Geraldine Mark is a research student at Mary Immaculate College, University of Limerick. Her main interests are in second language development and spoken language. She is co-author of *English Grammar Today* (2011, Cambridge University Press, with Ronald Carter, Michael McCarthy and Anne O'Keeffe) and co-principal researcher (with Anne O'Keeffe) of the English Grammar Profile, an online resource profiling L2 grammar development.

Eva Maria Martika is a PhD candidate in linguistic anthropology at the University of Toronto, supervised by Jack Sidnell. She has an MR in sociolinguistics and an MA in applied linguistics from the University of Essex. Her primary research interests are in conversation analysis and sociolinguistics. Her current research combines ethnography and conversation analysis to explore language socialisation practices in parent–child interactions.

Tony McEnery is Distinguished Professor of English Language and Linguistics at Lancaster University. He is the author of many books and papers on corpus linguistics, including *Corpus Linguistics: Method, Theory and Practice* (with Andrew Hardie, CUP, 2011). He was founding director of the ESRC Corpus Approaches to Social Science (CASS)

Centre, which was awarded the Queen's Anniversary Prize for its work on corpus linguistics in 2015.

Mark McGlashan is Lecturer in English Language within the School of English at Birmingham City University. His research interests predominantly centre on corpus-based (critical) discourse studies and the analysis of a wide range of social issues typically relating to nationalism, racism, sexism and homophobia.

Emma Moreton is a Lecturer in Applied Linguistics in the Department of English at The University of Liverpool. Her current research uses corpus and computational methods of analysis to examine the language of historical correspondence data, with a particular focus on nineteenth-century letters of migration.

Anne O'Keeffe is Senior Lecturer at Mary Immaculate College, University of Limerick. Her research includes co-authoring *Investigating Media Discourse* (2006, Routledge), *From Corpus to Classroom* (2007, Cambridge University Press), *English Grammar Today* (2011, Cambridge University Press), *Introducing Pragmatics in Use* (1st ed. 2011, 2nd ed. 2020, Routledge). She is co-editor of the *Routledge Handbook of Corpus Linguistics* (1st ed. 2010, 2nd ed. forthcoming with Michael McCarthy). She was co-principal researcher (with Geraldine Mark) of the English Grammar Profile, an online resource profiling L2 grammar development, and is co-editor of two Routledge book series: Routledge Corpus Linguistic Guides and Routledge Applied Corpus Series.

Laura L. Paterson is a Lecturer in English Language and Applied Linguistics at The Open University, UK. Her work involves using corpus linguistics and critical discourse analysis to focus on UK poverty and audience response to poverty porn. She has also written on discourses of marriage and epicene pronouns.

Peter Petr   is Research Professor of English and General Linguistics at the University of Antwerp, Belgium. Combining linguistics, history and cognitive science with data-oriented methodologies, his research focuses on how grammar evolves within individual cognition and how macro-properties emerge from individual interactions at the community level. His publications include *Constructions and Environments* (2014, Oxford University Press) and *Sociocultural Dimensions of Lexis and Text in the History of English* (2018, John Benjamins, co-edited with Hubert Cuyckens and Frauke D'hoedt). He has also co-developed the large-scale corpus *Early Modern Multiloquent Authors* (EMMA).

Mario Saraceni is Reader in English Language and Linguistics at the University of Portsmouth. His main research interests are in the politics and ideology of English as a global language. His publications include *English in the World* (edited with Rani Rubdy, 2006), *The Relocation of English* (2010) and *World Englishes: A Critical Analysis* (2015).

Elena Semino is Professor of Linguistics and Verbal Art in the Department of Linguistics and English Language at Lancaster University and director of the ESRC Centre for Corpus Approaches to Social Science. She holds a visiting professorship at the University of Fuzhou in China. She has co-authored four monographs, including *Metaphor in Discourse* (Cambridge University Press, 2008) and *Figurative Language, Genre and Register* (Cambridge University Press, 2013, with Alice Deignan and Jeannette Littlemore). She

is co-editor of the *Routledge Handbook of Metaphor and Language* (2017, with Zsófia Demjén).

Jack Sidnell is a linguistic anthropologist who earned a PhD from the University of Toronto in 1998 for a dissertation supervised by Hy van Luong. He has conducted ethnographic and sociolinguistic research in Canada, in the Caribbean and in Vietnam. He has published papers on conversation analysis, linguistic pragmatics, sociolinguistics and linguistic anthropology. He is the author or co-author of three books and the editor or co-editor of several more, including the *Handbook of Conversation Analysis* (2012). He is currently a professor of anthropology at the University of Toronto.

Paul Spence is Senior Lecturer in Digital Humanities at King's College London and has an educational background in Spanish and Spanish American studies. His research currently focuses on digital publishing, global perspectives on digital scholarship and interactions between modern languages and digital culture. He leads the 'Digital Mediations' strand on the Language Acts and World-Making Project (<https://languageacts.org/>).

Irma Taavitsainen is Professor Emerita of English Philology at the University of Helsinki. Her current research on historical pragmatics focuses on genre developments and speech acts, corpus linguistics and the evolution of scientific thought styles in medical writing.

Caroline Tagg is Senior Lecturer in Applied Linguistics at The Open University, UK, with an interest in how social media constrains and enables different forms of communication. Her research rests on the understanding that digital communication practices are embedded into individuals' wider social, economic and political lives, with implications for digital literacies education.

Jonathan Tallant is Professor of Philosophy at the University of Nottingham, where he's worked since 2007. Before coming to Nottingham he worked at the University of Leeds and completed his PhD at Durham University, working with Jonathan Lowe. His main areas of research interest are in metaphysics, where he's published extensively on the philosophy of time. Jonathan also has research interests in other areas including the philosophy of science and the philosophy of trust.

Riki Thompson is Associate Professor of Rhetoric and Writing Studies at the University of Washington Tacoma. Her current research employs multimodal critical discourse studies to examine storytelling, identity construction and belonging through digitally mediated spaces.

Alexandra Trachsel (PhD, habil.) is a Senior Lecturer at the University of Hamburg. Her main fields of interest are ancient Greek literature and the transmission of knowledge in antiquity, but since 2008 she is also involved in digital humanities and co-organised the panel 'Rethinking Text Reuse as Digital Classicists' at the DH2014 conference in Lausanne. Furthermore, she participated at the Digital Classicist Seminar in London (in 2009) and at the Digital Classicist Seminar in Berlin (in 2013). She also spent a Marie Curie Intra-European Fellowship at the Department of Digital Humanities at King's College London (2012–2013).

Freek Van de Velde is a Research Professor at KU Leuven, focusing on language variation and change, historical linguistics, quantitative linguistics and evolutionary linguistics. Previously he has worked on several postdoctoral projects funded by Lessius Research Fund, Leuven BOF Research Fund and FWO and taught a broad range of courses, mainly in the field of language variation and change and historical linguistics at KU Leuven, University of Gent, Universität zu Köln and the Westfälische Wilhelms-Universität Münster.

Piia Varis is an Associate Professor at the Department of Culture Studies and deputy director of Babylon, Centre for the Study of Superdiversity at Tilburg University, the Netherlands. Her research interests include digital culture, social media and the role of digital media and technologies in knowledge production.

Viola Wiegand is a Research Fellow on the CLiC Dickens project at the University of Birmingham. Her main research interest is the study of textual patterns with tools and frameworks from corpus linguistics, stylistics and discourse analysis. She is assistant editor of the *International Journal of Corpus Linguistics* (John Benjamins) and has co-edited *Corpus Linguistics, Context and Culture* with Michaela Mahlberg (De Gruyter, 2019).

John Williams is Senior Lecturer in English Language and Linguistics at the University of Portsmouth, specialising in corpus linguistics and lexicology. Earlier in his career, he worked as a lexicographer on various well-known language reference publications, notably the innovative COBUILD series. He has also co-edited a handbook for international online teaching in the humanities (Teaching Across Frontiers, 2001).

David Wright is a forensic linguist and Senior Lecturer at Nottingham Trent University. His research applies methods of corpus linguistics and discourse analysis in forensic contexts and aims to help improve the delivery of justice using language analysis. His research spans across a range of intersections between language and the law and justice; language in crime and evidence; and discourses of abuse, harassment and discrimination.

Acknowledgements

As an area of research, the intersection between English language and digital humanities is relatively new, and we are grateful to the large number of people who have encouraged us to explore this area in more detail, written chapters, provided feedback on structure and content and offered valuable reviews on individual chapters. In addition to chapter authors who have so cogently embraced the English language–digital humanities interface, we would like to thank the following colleagues for their input, feedback and detailed reviews: Robert Bayley, Lou Burnard, Ronald Carter, Alan Chamberline, Patrick Finglass, Tess Fitzpatrick, Kevin Harvey, Spencer Hazel, Judith Jesch, Deborah Keller-Cohen, Mike McCarthy, Jeremy Morely, Ruth Page, Amanda Potts, Graham Stevens, Peter Stockwell, Peter Stokes, Melissa Terras, Karin Tusting, Rachelle Vessey, Keir Waddington and Sara Whiteley. Special thanks also to our advisory board, especially for their early help in shaping the content and structure of this volume: Wendy Anderson, Susan Conrad, Rodney Jones and Martin Wynne.

Getting to the final stages of submitting the handbook has taken much longer than we had anticipated, and we are grateful to colleagues at Routledge for their patience, sound advice and support along the way. Our thanks to Nadia Seemungal-Owen for commissioning the handbook and to Elizabeth Cox, Helen Tredget and Adam Woods for their outstanding editorial support and guidance. Thanks also to Val Durow for her invaluable help with proofreading the whole volume and to Autumn Spalding for her expert advice throughout the prepress stage of this publication.

Every effort has been made to trace the copyright holders for material in the individual chapters of this volume. Where this has not been possible, the publishers would be glad to have any omissions brought to their attention so that these can be corrected in future editions. We would like to thank *Mumsnet* for permission to use the *Mumsnet* logo and excerpts from the *Mumsnet* Talk used in Chapter 4.

The idea for this handbook was born in a discussion we had with the late Ron Carter who supervised both our PhD theses and later became a close colleague, mentor and friend. He offered encouragement and advice throughout the process, and we are sad that he is no longer here to see the final version in print. We dedicate this book to him.

Svenja Adolphs and Dawn Knight
April 2020



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

English language and the digital humanities

Svenja Adolphs and Dawn Knight

As a humanities discipline, the empirical study of the English language can be deemed to be amongst the ‘early adopters’ of the digital. With its exploration of language patterning across diverse language corpora, it has embraced the affordances of computer technology early on and reaped the benefits by developing more nuanced descriptions of the language, and improved applications based on those descriptions. It has also maintained a close relationship with other disciplines across the broad spectrum of the academy, especially with other sub-disciplines of the humanities and the social sciences. At the same time, the wider humanities have embraced digital methods and approaches for analysing, preserving, linking and ‘fusing’ data in their own respective fields, ranging from large cultural datasets to carefully annotated multimodal episodes of discourse. This handbook serves as a reference point for key developments related to the ways in which the digital turn has shaped the study of the English language and of how the resulting methodological approaches have permeated other disciplines within the humanities. It draws on modern linguistics and discourse analysis for its analytical methods and applies these approaches to the exploration and theorisation of issues within arts and the humanities. The handbook is intended for a broad audience and doesn’t assume any prior knowledge of the field. It should be equally useful to newcomers in English language research and to those interested in how language can be analysed through digital interfaces and in digitally supported ways across other disciplines in the humanities, and alongside other types of data and methods used in those disciplines.

The digital turn in English language research and in the humanities

The study of the English language has benefitted significantly from advances in computer technology since the 1960s, when the first language corpora became more widely available. Preservation and techniques for sharing language data were afforded an increasingly prominent place in the research community, and new insights into the frequencies of words and phrases across different contexts and different language varieties

were generated on the back of ever-growing datasets. Corpus creation and sharing, corpus-based lexical studies and machine translation were often categorised under the broader heading of humanities computing. Digital search and manipulation of data turned out to be not only relevant to quantitative approaches but soon also found its way into more discourse-based modes of analysis. Whilst the vast majority of corpus-based research has tended to focus on written material, there are some notable exceptions where spoken language was recorded and transcribed in a digital format, making these datasets available for annotation, search and sharing. The early twenty-first century marked the arrival of multimodal corpora and the study of gesture and prosody alongside video capture and textual transcripts of spoken interactions at scale. Much has been written about the way in which the digital approach to the study of English has impacted on language descriptions and theories, and parts one and two of this volume give both an overview of corpus data as evidence and a range of in-depth case studies which illustrate how the various sub-fields of linguistics have evolved alongside digital technologies for language research and have started to explore ‘born-digital’ language in its own right.

The digital exploration of samples of language shares certain aspects with the digital humanities. As Jensen (2014) points out, early efforts in the area of digital humanities were often met with scepticism by those holding more traditional views, and there are certain parallels with the early developments in using digital methods and empirical evidence in linguistics: the focus on empiricism in language research presented a challenge to the theoretical approaches of Chomskyan linguistics that were widely used at the time, relying on speaker intuition as evidence rather than on language in use. However, despite these parallels there are also key differences when it comes to using digital tools in the study of the English language and the digital humanities. They include the types of data that are being used and the methods that are being applied to the interrogation of the data. This volume covers both quantitative and qualitative methods, both of which we see as being key to modern linguistics and digital humanities, and the key topics in part three of this volume discuss approaches to data linkage and exploration. As such, we believe that this handbook illustrates the mutual benefits that emerge from the study of the English language in and through digital humanities contexts.

The chapters in this volume adopt different definitions and conceptualisations of what we mean by digital humanities and how the term is related to their respective area of study. The term is sometimes used as a proxy for a more inclusive and interdisciplinary approach when compared to more narrow definitions, such as humanities computing (see also Terras 2016; Kirschenbaum 2014). The inclusive outlook of digital humanities is thus aligned with the growth of other emerging, and re-emerging, areas in the digital research space over the past two decades, such as e-science, e-social science, data science and artificial intelligence, and is likely to contribute as much to the fostering of new collaborations as it does to addressing long-standing issues in the arts and humanities and beyond. Digital humanities research is gathering pace in permeating and transcending more established sub-disciplines that deal with humanities issues. Regardless of whether digital humanities becomes completely embedded in research practice or continues to grow as a field of research in its own right, the remit of using ‘computational methods that are trying to understand what it means to be human, in

both our past and present society’ (Terras 2016: 1637-1638) is likely to gain in importance across all disciplines and applied contexts. And just like any area of research that relies heavily on digital approaches and data, it will have to embrace the legal and ethical issues that are a key part of this territory and should be well placed to contribute to this debate.

The scope of this volume

This handbook is broadly divided into three parts. The first part deals with different types of linguistic data, either rendered into digital format or ‘born-digital’ (Chapters 2–6). The chapters in this section give an overview of the sources that scholars working with the English language tend to draw on, regardless of whether this is intended to advance our understanding of the English language as such or of applications to other disciplines as outlined in part three. The chapters also include a general overview of issues of corpus construction, including the principled selection of samples, breadth and size of corpora, metadata, capture, storage and processing of multimodal data, as well as issues around ethics and consent.

The second part covers a range of approaches within modern linguistics that rely on digital data and methods in their analysis (Chapters 7–20). The chapters illustrate how some of the more traditional linguistic approaches and frameworks have evolved alongside the use of corpora and corpus linguistic techniques. Many of the approaches in this part have undergone a major conceptual and analytical transition when they started drawing on corpus data, and the chapters highlight the various angles of this transition.

The third part of the volume moves into the wider field of humanities research, exploring the use of linguistic methods in digital humanities contexts (Chapters 21–30). Individual chapters highlight some of the key issues in their respective disciplines and explore ways in which methods and data from modern linguistics can be brought to bear on those issues. This section covers a wide range of disciplines and data sources used in the humanities and at the interface to the social sciences, and illustrates interesting ways in which linguistic and other data sources and theories can be aligned and studied alongside one another.

This handbook focuses on different varieties of the *English* language and the digital humanities, and we are aware of the limitations of exploring and theorising issues in the digital humanities through a lens that is mainly restricted to one language. However, the work in this volume serves to illustrate the exploration of the interface between digitally inflected research of the English language, often situated in multilingual contexts, and the digital humanities within a larger context of all languages.

While Jensen sees corpus linguistics as located at the ‘fringes of contemporary DH, which is itself currently on the fringes of the humanities’ (2014: 131), especially the chapters in part three of this handbook show ample evidence that corpus linguistics, discourse analysis, the humanities and the digital humanities are becoming more interconnected. We hope that this volume will make a useful contribution to this area of research which we believe will grow and diversify alongside digital advances, both in terms of the plurality of approaches and the languages that are at the heart of this interface.

References

- Jensen, K.E. (2014). Linguistics and the digital humanities: (Computational) corpus linguistics. *MedieKultur* 57: 115–134.
- Kirschenbaum, M. (2014). What is ‘Digital Humanities,’ and why are they saying such terrible things about it? *Differences* 25(1): 46–63.
- Terras, M. (2016). A decade in digital humanities. *Journal of Siberian Federal University. Humanities & Social Sciences* 9: 1637–1650.

Spoken corpora

Karin Aijmer

Within a very short period, linguists have acquired new techniques of observation. The situation is similar to the period immediately following the invention of the microscope and the telescope, which suddenly allowed scientists to observe things that had never been seen before. The combination of computers, software and large corpora has already allowed linguists to see phenomena and discover patterns which were not previously suspected. To that extent, the heuristic power of corpus methods is no longer in doubt (Stubbs 1996: 231).

Introduction

The humanities can be defined as the academic disciplines traditionally concerned with seeking ‘to understand the human experience, from individuals to entire cultures, engaging in the discovery, preservation, and communication of the past and present record to enable a deeper understanding of contemporary society’ (Terras 2016: 1637). The digital humanities can then be defined as the discipline concerned with what it means to be a human being and our place in human society and culture. The definition in *Wikipedia* provides a good idea of the methods and scope of this new discipline:

The digital humanities, also known as humanities computing, is a field of study, research, teaching, and invention concerned with the intersection of computing and the disciplines of the humanities. It is methodological by nature and interdisciplinary in scope. It involves investigation, analysis, synthesis and presentation of information in electronic form. It studies how these media affect the disciplines in which they are used, and what these disciplines have to contribute to our knowledge of computing.

Background

Scholarship in the humanities is traditionally about decoding, understanding and critiquing texts of the past and the present and reflecting on what they tell us about human culture. The

digital landscape is now changing, and computer-based methods are also coming to play an important role in the study of literary, cultural and historical texts, including the arts, music, film and theatre.

However, digital humanities and corpus linguistics can be said to represent different traditions and histories sharing an interest in how computer-associated tools can be used in the study of texts. Corpus linguistics is concerned with collecting and designing corpora and using corpora and corpus linguistic methods to study language. Digital humanities research is about the development and exploitation of databases or data repositories where transcribing the material and using special mark-up are important. Such a project is the critical edition of Jeremy Bentham's writings, for example (Terras 2011).

Linguists have now started to fully exploit corpus methods and the resources of spoken corpora, with a flourishing of research on spoken grammar as a result. As will be further discussed, corpus linguistic methods are also applied to fields which are interesting not only to linguistics but to practitioners in fields such as education and second language acquisition, healthcare communication, storytelling and sociology.

In this study I will address the problems of compiling a spoken corpus, types of spoken corpora, annotation and mark-up, as well as give examples of the contribution of spoken corpora to research in a large number of applied enterprises. The chapter is structured as follows. The chapter begins with a methodological discussion of issues in designing spoken corpora. It is followed by a section discussing topics in variational discourse analysis using corpora. Current contributions and research are illustrated by the use of spoken corpora in sociolinguistic studies of variation and change. The sample analysis demonstrates the use of spoken corpora in pragmatics, discourse and in (critical) discourse analysis. The chapter ends with a discussion of future developments envisaged by research using spoken corpora.

Main research methods

The aim of this section is to back up the argument that a field such as discourse humanities can benefit from using and collecting spoken corpora for particular problem areas. To begin with, I will discuss challenges in compiling a spoken corpus with regard to the representativeness and the size of the corpus, accessibility, mark-up and transcription, annotating the data with pragmatic information.

Compiling and designing a representative spoken corpus

The collection of spoken corpus data starts by recording speakers and then transcribing the data. The priority when selecting the data is that the corpus should be representative. However, especially if the aim is to collect a general spoken corpus, the balancing of the total number of speakers with regard to demographic variables may be problematic. Ideally, the contributors should be recruited in equal numbers of male and female speakers, represent different age categories, etc. This may be difficult for practical reasons. The recent Spoken British National Corpus (BNC) 2014, for example, therefore uses a more opportunistic approach where the data represent nothing more or less than it was possible to gather for a specific task (McEnery and Hardie 2012: 11).

Different recruitment methods may be used. A method which is also becoming common in digital humanities is to recruit participants through crowd sourcing, that is, asking for help from the general public rather than asking contributors privately. The aim to represent

authentic spontaneous language where speakers are unaware of being recorded has been difficult to realise because of ethical principles. Participants in the Spoken BNC 2014 project were instructed to use portable smartphones to record conversations with their friends or family. When recording the contributors, it is also important to consider their privacy. For example, it is necessary to have the contributors' previous permission in order to record them. As with any case of non-surreptitious data, it follows that the observer's paradox cannot be avoided. If the speakers know they are being recorded, they are likely to adapt their speech to the situation.

Size of the corpus

Recording the conversation and transcribing the data is a time-consuming process which has delayed the advancement of spoken corpora. However, the transcription process may be simplified by access to a transcription software program (such as ELAN or Praat). Moreover, the corpora contain concordancers and can be linked to a special querying system such as the Corpus Query Processor (CQP) workbench or the Sketch Engine (www.SketchEngine.co.uk/) to search for information from the corpus.

Corpora are increasing in size. The *London-Lund Corpus of Spoken English (LLC)*, recorded in the 1960s and 1970s, contained half a million words (Svartvik 1990). This is not considered big by today's standards, and the LLC has been succeeded by several ambitious projects to collect large amounts of spoken data. To give an example, the *BNC* (www.natcorp.ox.ac.uk/) consists of 10 million words of spoken language, and the *Bank of English* (<http://www.cqpweb.bham.ac.uk/>) (associated with the 'Birmingham school' of corpus linguistics) has a spoken component of about 4.5 million words. Other 'big' spoken corpora are the *Cambridge and Nottingham Corpus of Discourse in English (CANCODE)* (McCarthy 1998), with a size of about 5 million words, and the *Longman Spoken and Written English Corpus (LSWE)*, with a spoken component of roughly 4 million words (see Biber et al. 1999). Most spoken corpora represent British English. An exception is the *Santa Barbara Corpus of Spoken American English (SBCSAE)*, which consists of recordings from different regions of the United States (249,000 words). The collection of a new BNC (*the Spoken BNC 2014*) representing English as spoken in the twenty-first century was completed in 2017. It matches the size of the spoken component of BNC 1994 (10 million words), which makes it possible to use the two corpora in parallel to describe short-time changes in the language using the available sociolinguistic metadata concerning age, gender, social class and region of the speakers. However, 'size isn't everything' (McCarthy and Carter 2001), and both large and small spoken corpora have their advantages. Big corpora are suitable for the quantitative analysis of large amounts of data. It is obvious we also need smaller specialised corpora compiled for special purposes or to investigate a particular text type, which allows researchers to make a more qualitative analysis of the data in their social context.

The MICASE Corpus (the Michigan Corpus of Academic Spoken English), which will be further discussed, was created to allow speakers to study academic speech and has inspired a British English equivalent (BASE 'the British Academic Spoken English Corpus'). The Narrative Corpus (NC) is a specialised corpus of naturally occurring narratives extracted from the demographically sampled sub-corpus of the BNC. The annotation includes features which are of interest if one wants to study narratives such as social information about the speaker, the textual components of the narrative, participant roles and quotatives (Rühlemann and O'Donnell 2012). There are also specialised corpora dealing with a particular

topic, such as the ‘fox hunting’ corpus compiled by Baker (2006) and further discussed in a later section, where I also discuss other specialised corpora used to study social or political issues critically.

Transcribing data from general corpora

The compilation of spoken corpora faces special challenges. A large number of features need to be annotated in the transcribed text, reflecting what Biber et al. (1999: 1048) refer to as ‘normal dysfluency’. Adolphs and Carter (2013: 12) mention the following key features that one would want to be marked up: ‘who is speaking when and where to whom; interruptions, overlaps, back channels, hesitations, pauses, and laughter as they occur in the discourse; as well as some distinct pronunciation and prosodic variations’. Inevitably researchers may select different (and fewer or more) features, and there will be many different ways of transcribing the spoken data. The British component of the International Corpus of English (ICE-GB) typically marks features such as speech units, incomplete words, overlapping speech, pauses, uncertain hearing and types of vocal noise. Some examples of the annotation from the ICE Corpus follow.

Incomplete word

<.> ... </.>

In the following example, the speaker interrupts herself. The beginning of the interruption is marked by <.> and the completion by </.>

our <.>we</.> weather ranged from days like this

The annotation is used when a word is thought to be incomplete.

Pause

<,> <,,>

Pauses can be of different lengths. They are judged impressionistically by the speaker.

Type of vocal noise

yeah you forget about that <laughs>anyway</laughs>

Other paralinguistic features are <whispering>, <mimicking American accent> or non-verbal sounds <nv> breathe<nv>, <nv> burp<nv> (examples from the *Bergen Corpus of London Teenage Language* [COLT]).

Another issue concerns how we should describe the structural units of spoken grammar corresponding to sentences or clauses in written grammar. Spoken language is dynamic in the sense that it is produced ‘here and now’, which has the effect that speakers are forced to adapt their speech to the constraints enforced by the ‘limited planning ahead principle’ (Biber et al. 1999: 1067). A characteristic feature of spoken language is that it is made up of chunks of different sizes corresponding to planning or idea units.

By way of illustration, an extract is shown from the LLC. The speakers taking turns in the conversation are identified as A> and B>:

- A> well 'eight * Cr\ouch End##
 B> *I'll have a look on the* t\ube 'map# I'll remember the t\ube 'station# -I
 ((don't)) know 'where Crouch \End 'is#
 (– laughs) but I might be [m] I ^think I –
 can't re'member the tube *st\ation##
 A> (*-* – laughs) – ^n\o# -
 B> “wasn't very far aw\ay#.
 it might have 'been 'Belsize P\ark# oh well !that's 'where his m\other l\ives#. >
 ^mother 'lives at 'Bel'size P\ark#

In the LLC the structural units are assumed to be units of intonation (tone units) marked by boundaries (#) in the transcription and reflecting the fact that speakers organise their messages as comprehensible ‘chunks’ of different lengths.

The LLC also marks intonation [/ \] and accent ['] as shown in the example. [*] marks overlapping speech. Pauses of different lengths are marked by a full stop or by dashes [. – –].

The conventions used for transcription and annotation are specific to the corpus that is being constructed and the research questions that will be asked of the corpus. The Lancaster/IBM Spoken English Corpus (SEC) consists mainly of prepared or semi-prepared speech and was used to construct a model for speech synthesis by computer (Knowles 1996: 1). It uses the same transcription system as the LLC but in a simplified form. The boundaries between tone units are marked, but the number of symbols for marking other prosodic phenomena is kept to a minimum. Only accented syllables are annotated with an indication of pitch movements, while paralinguistic features such as tempo, loudness and voice quality are not marked at all. The study of prosody has so far been fairly limited in spoken corpus studies. Existing research in this area shows how prosodic practices are linked to lexical items or constructions (Aijmer 1996) and, more broadly, how syntactic, prosodic and contextual features interact to describe the meaning of pragmatic expressions. Wichmann (2004) used spoken data from the ICE-GB to study the intonation of requests with *please*. She first searched for all occurrences of *please* in the corpus and categorised the occurrences according to formal, functional and contextual criteria. In the next stage, intonation contours were assigned to *please* based on the audio recordings. The results showed that in situations where *please* could be interpreted as speaker-oriented, the request ended in a fall. Where *please* was used to appeal to the hearer, the request ended in a rise, conveying a greater sense of obligation.

However, richly transcribed corpora such as the LLC or the SEC are now unusual. An exception is the *Hong Kong Corpus of Spoken English (HKCSE)*, which exists in both an orthographic and a prosodically transcribed version (Cheng et al. 2008). On the other hand, many corpora (such as the spoken component of the corpora collected under the umbrella of the ICE Corpus [www.ucl.ac.uk/english-usage/projects/ice.htm]) give access to the audio files accompanying the recordings, which can be used for the prosodic analysis. Through a process referred to as time alignment it is also possible to hear the portion of the recording which matches the item searched for (McEnery and Hardie 2012: 4). This technique, which is useful both for transcription and for the subsequent analysis of the data, is now also used for annotating multimodal corpora. It is also possible to use the software program Praat to analyse speech from corpora in phonetics (www.fon.hum.uva.nl/praat/).

Transcribing data from specialised corpora

Specialised corpora have special needs and goals, which may provide transcription problems. An example is child language, where the data from the recordings may be difficult to interpret. In this case we also need a ‘rich’ transcription which incorporates information that will be of interest to researchers on child language. This is illustrated by the Child Language Data Exchange System (CHILDES), a large database collected from conversations between children and their playmates or caretakers (Lu 2010).

The conversations were transcribed at different levels. Each utterance was followed by a line containing morphological analysis. Any physical action accompanying the utterance was also marked. In addition, there were links between the digital and video recordings, and the transcriptions were compatible with software programs such as ELAN and Praat. The corpus is used with the software program CLAN containing computational tools designed to analyse the data transcribed. A special file is available which provides detailed demographic, dialectological and psychometric data about the children, as well as a description of the context.

Regional and dialectal corpora provide another example where we need a fine-grained transcription. The vernacular Newcastle Electronic Corpus of Tyneside English (NECTE) was transcribed on different levels, including information about Standard English orthography, segmental phonology, prosodic and grammatical features. In addition, data was provided about the speakers (Kretschmar et al. 2006).

Pragmatic annotation of spoken corpora

We have seen a shift in corpus linguistics from the ‘big data’ paradigm to an emphasis on ‘rich’ corpus data (cf Wallis 2014: 13). Spoken corpora need to be annotated in a way that allows the researcher to identify linguistic elements having a pragmatic or discourse function, such as speech acts or discourse markers. However, annotation schemes are still very much in their infancy. As Weisser (2015: 84) points out, ‘any type of linguistic annotation is a highly complex and interpretive process, but none more so than pragmatic annotation’. Moreover, it is an aim to be able to automate the annotation process as much as possible to increase the speed and to ensure that it is systematic and consistent.

Weisser emphasises the importance of using some form of Extensible Markup Language (XML) to mark up pragmatic units (Weisser 2015: 86). Perhaps the most obvious functional units to be marked up in this way are those representing different levels in the hierarchical discourse structure, such as transactions, turns, backchannels and adjacency pairs (exchanges). See further Weisser (2015) and Kirk and Andersen (2016) on the different annotation schemes enriching corpora with discourse information.

It is also important to annotate speech acts, discourse markers and formulaic expressions such as *thanks* so that one can study them quantitatively from a form-to-function perspective. Kirk and Kallen (see e.g. Kallen and Kirk 2012; Kirk 2016) found it desirable to pragmatically annotate quotation, speech acts, discourse markers and vocatives in the *Systems of Pragmatic Annotation in the Spoken Component of ICE-Ireland (SPICE-Ireland Corpus)* in order to facilitate the comparison of such expressions across the corpus. Pragmatic annotation of discourse elements, however, is difficult to do automatically because the same form may represent many different grammatical categories, and other cues, such as position in the utterance, may not be reliable. Thus, *now* may be both a temporal adverb

and a discourse marker, and there are examples where *now* is analysed in both ways. In the following example (adapted from Kallen and Kirk 2012: 47), *now* is a discourse marker drawing attention to a new discourse frame in quoted speech. It contrasts with the temporal *now* in the same clause:

```
<P2A-022$A> <#> <rep> And I said+ <quote>
now* where do I go now </quote> </rep>
(<rep> indicates a representative speech act)
```

Ongoing work in this area also includes developing schemes for annotating hesitation markers such as ‘um’ and backchannels or listener responses in the discourse (Weisser 2015). (See further Rühlemann and Aijmer 2015: 11.)

Spoken genres

Spoken corpora should ideally cover a large number of different genres (or text types). This makes it possible to study the variation or changes which are dependent on aspects of the context, such as whether the speakers are involved in conversation, participating in a broadcast discussion or appearing in a classroom situation. The LLC, for example, contains texts from face-to-face conversation, telephone conversation, discussion, public speech and prepared speech, and the British component of the ICE-Corpus mirrors this division into genres, as seen next:

- **Private (dialogue)**
 - Direct conversation
 - Telephone calls
- **Public (dialogue)**
 - Classroom lessons
 - Broadcast discussions
 - Broadcast interviews
 - Parliamentary debates
 - Legal cross-examinations
 - Business transactions
- **Unscripted monologues**
 - Spontaneous commentaries
 - Unscripted speeches
 - Demonstrations
 - Legal presentations
- **Mixed**
 - Broadcast news
- **Scripted monologues**
 - Broadcast talks
 - Non-broadcast speeches

The ‘context-governed’ component of the BNC contains business meetings, parliamentary debates, public broadcasts, etc., and the demographically sampled component of the corpus records speakers of different ages, regions and social classes. However, there is no fixed number of genres or ‘activities’ in spoken language, and general corpora usually contain mostly informal conversation. Moreover, what constitutes naturally occurring spoken data can be interpreted in different ways. In the Corpus of Contemporary American English (COCA) the data are described as being naturally occurring and not scripted. However, the spoken data are restricted to TV and radio programmes since these varieties are easier to extract and are publicly available (cf Davies 2010: 453).

We also need to address the question of the balance between different genres. A representative and balanced corpus can, for example, aid in differentiating between linguistic choices relative to whether the situation is an informal conversation or a classroom lesson. In the BNC this was achieved by distinguishing between a demographic part and a context-governed part representing many different situations. The CANCODE can also be considered a balanced corpus, although the categories proposed to represent the genres of casual conversation are different, reflecting the fact that the corpus design was more fluid to begin with.

The categories introduced reflect the degree of familiarity between the speakers, such as intimate, professional, transactional and pedagogic, and ‘typical goal types’, such as information provision (‘unidirectional discourse’) and collaborative ideas (‘bidirectional discourse’) (Adolphs and Carter 2013: 39). Conversations that take place between members of the same family have been classified as intimate, while the sociocultural category involves interactions between friends. The professional category covers discourse that is related to professional interactions, and transactional categories would be exemplified by the interaction between a waiter and a customer at a restaurant.

Critical issues and topics

Spoken language corpora open up new perspectives on studying variation in language use depending on the speech situation, society and culture. One hypothesis is that there are differences when we move from one regional variety of English to another or compare native and non-native speakers. Corpus-based studies on variation have in common that they compare the frequency of a pragmatically interesting feature in one corpus with the same feature in another.

Variation across national languages

Studies in variational discourse analysis may involve two different languages. In order to study variation across languages, we need access to comparable corpora. Stenström (2006) used the Spanish corpus *El Corpus de Lenguaje Adolescente de Madrid (COLAm)* (<http://colam.org/publicaciones/corpuslenguajeadoles.htm>) and the *COLT Corpus* (Stenström et al. 2002) to compare the Spanish discourse markers *o sea* and *pues* with their close English correspondences *cos* and *well*. What this and similar studies show is that there is not a direct correspondence between elements in the compared languages. *Pues* may, for example, have several different correspondences in English (including *cos* ‘because’ and the discourse marker *well*) and, in many cases, it had no particular correspondence at all.

Corpus linguists are also interested in comparing phenomena cross-linguistically through the use of corpora on a common topic or representing the same genre. White et al. (2007) used three parallel corpora of political debate (Swedish, Dutch, British) to study how *of*

course and its correspondences in the two other languages were used to express ‘taken-for-grantedness’. It was shown that the various markers were used for the same purpose to mark that something was taken for granted in the debates. However, there were also differences between the languages, indicating that ideology may play a role and that the politicians have different ideas about a political debate.

English has had a phenomenal spread across the globe. As a result, the notion ‘English’ now also refers to new ‘Englishes’, such as Singapore English, Philippine English and Hong Kong English. The ICE Corpus (<http://ice-corpora.net/ice/>) is an ongoing project with the aim of compiling a corpus of varieties of English. The sub-corpora have been designed in the same way, enabling comparisons across the varieties either in informal conversation or in more formal varieties of spoken English. Each sub-corpus consists of or will consist of 1 million words of spoken and written English (produced after 1989). The present ICE corpora (containing a spoken component) include Canada, East Africa, Great Britain, Hong Kong, India, Ireland, Jamaica, New Zealand and the Philippines. The corpora can be used to describe the characteristic features of the new varieties synchronically and, more broadly, look for models explaining the spread of English and the formation of new varieties.

Aijmer (2015) studied *actually* on the basis of its occurrence in four sub-corpora of ICE representing Britain, New Zealand, Singapore and Hong Kong. It was shown that there were large differences in the frequency of *actually* between the varieties, with *actually* having the highest frequency in ICE-HK and being least frequent in ICE-NZ. There were also some tendencies to positional and functional specialisation. In Singapore English *actually* was mainly used in the leftmost ‘outside -the -clause’ position, with the function of emphasising the speaker’s perspective. In Hong Kong English *actually* was typically used together with *but* to oppose a preceding claim.

Variation across time

Spoken corpora have now been in existence for such a long time that we can make comparisons between spoken data from different periods. The LLC, compiled in the 1960s and 1970s, was one of the earliest spoken corpora. The data from the LLC can now be compared with authentic conversation recorded later in the British component of the ICE Corpus. The *Diachronic Corpus of Present-Day Spoken English (DCPSE)* (www.ucl.ac.uk/english-usage/projects/dcpse/) is a ‘comparative’ corpus, which makes it possible to contrast conversational texts in the two corpora and to study short-term changes – for example, in the area of discourse markers.

A project still ‘in the air’ is the updating of the original LLC (the *London-Lund Corpus of Spoken British English 2*) (<https://projekt.ht.lu.se/llc2>). However, a new Spoken BNC (*Spoken BNC 2014*) (<http://language-research.cambridge.org/spoken-british-national-corpus/>) has recently been released, which will be a resource for the sociolinguistic analysis of variation and change. As a result, short-time changes can now be studied by comparing the ‘traditional’ BNC from the 1990s with the more recent Spoken BNC 2014. The comparison shows, for example, that the rise of intensifying adverbials such as *really* and *so* in contemporary British English is striking, while *very* has become less frequent (Aijmer 2018).

Spoken corpora and sociolinguistics

There is abundant evidence that the social factors age, gender and social class have an effect on the language used in conversation. In line with this, spoken corpora generally include

some demographic data about the speakers. The LLC, for example, gives some social information about the speakers (age, gender, relationship to the hearer), and the BNC – in its demographic part – provides sociolinguistic metadata about the speakers. However, so far, corpus linguistic methodology has only had a limited role in sociolinguistics. One reason is that available corpora such as the BNC are not perfectly balanced for variables such as the age and gender of the speakers and therefore less suitable for traditional variationist sociolinguistics as represented by Labov (1966, 1972).

In contrast to the sociolinguistic variationist paradigm, the focus in corpus linguistics is on the function of a linguistic form and the factors determining the speaker's choice of the linguistic expression. Erman (1992) suggested that there are gender-specific differences in the use of the pragmatic expressions (*you know*, *you see*, *I mean*) on the basis of their frequencies in a sample of conversations from the LLC. An increasing number of scholars have recently used the BNC to study the sociolinguistics of pragmatic phenomena with regard to many different variables. Rühlemann (2007: 178) analysed the quotative *I says* with regard to the variables age, gender, social class and dialect of the users on the basis of the demographic part of the BNC. He found some evidence that it was used more by female than male speakers, that it was preferred by speakers between 35 and 44 years of age and that it was 'clearly not typical of upper-class usage' (Rühlemann 2007: 177). Moreover, *I says* was used most frequently by speakers in Northern England.

Beeching (2016) addresses the sociolinguistics of the common discourse markers *well*, *just*, *you know*, *sort of* and *I mean* in different functions on the basis of the BNC. The findings show, for example, that female speakers use all the markers more often than men (and that the differences are significant for *well*, *just*, *like* and *I mean*) (Beeching 2016: 216).

By including sociolinguistic factors, we get a more meaningful interpretation of how the markers are used in the interaction by real people.

Spoken corpora and the study of teenage language

Access to sociolinguistic data opens up the study of a wide range of age-related phenomena in spoken language (see Stenström et al. 2002). Teenagers are known to be linguistic innovators of language, which has prompted the compilation of special teenage corpora. The *Bergen Corpus of London Teenage Language (COLT)* is based on transcriptions of conversations between teenagers in the London area taking place in 1993. The corpus has a sociolinguistic design marking up information about the age, gender and social class of the participants. The recorded conversations amount to around 100 hours corresponding to about 500,000 words in the transcriptions. The availability of the corpus has put the study of swearing and 'bad language' on the map (see Stenström et al. 2002). It is also possible to investigate new uses of discourse markers. Andersen (2001) studied innovative uses of *like* (and the tag question *innit*) in COLT. The study showed that teenagers were in the forefront of people using *like* in new ways, which may signal an ongoing development resulting in grammatical change.

The Linguistic Innovators project was initiated with the purpose of characterising London English 'after COLT' considering the effect of massive multiculturalism in London on the spoken language. The empirical data are taken from the *Linguistic Innovator Corpus (LIC)* (Gabrielatos et al. 2010) built on interviews with London adolescents (including teenagers with a recent immigrant background ['non-Anglos']). In a case study using the corpus, Torgersen et al. (2011) compared discourse markers in COLT and LIC. On the basis of the results they argued, among other things, that a number of discourse markers such as *you get me* should be regarded as characteristic of Multicultural London English (MLE).

Another teenage-specific corpus, not based on London speakers, is the *Toronto Teen Corpus: 2002–2006 (TTC)*, used by Tagliamonte (2016) to study trendy discourse markers and intensifiers in the spoken language of Toronto teenagers. Tagliamonte found an overwhelming number of *like* in her corpus; moreover, the teenagers used many more tokens of *really* than their parents or grandparents.

Corpora and applied linguistics

Applied linguistics is ‘the academic field which connects knowledge about language to decision-making’ (Simpson 2010: 1). It is oriented to solving problems in the real world and makes insights drawn from language relevant to such decision-making. It is closely associated with language learning, teaching and language acquisition. However, the scope of applied linguistics is broad, and it also concerns matters such as language and identity, translation and contrastive linguistics, globalisation issues such as the spread of English, English as a lingua franca, health communication and media discourse. Applied linguistics also encompasses a special perspective on society as in critical discourse. The data used are typically empirical, and corpora have been an important resource in many areas of applied linguistics.

Spoken corpora of healthcare communication

Corpora and corpus linguistic methods have an important role in studying healthcare communication. The *Nottingham Healthcare Communication Corpus* is specialised in healthcare language. It takes into account a number of interaction types, for example, the interaction between healthcare practitioners such as nurses and doctors and, on the other hand, patients. Adolphs et al. (2004) used a sub-corpus consisting of transcribed telephone conversations between callers and health advisers and proceeded with a corpus linguistic investigation of the data using frequency lists and key word analysis. The analysis of the words which were most salient (‘key words’) in comparison with a corpus of general English suggested that the advisers used a style characterised by personal pronouns, vague language and softening words. The results provide medical educators with information about linguistic features characteristic of healthcare practices, which could improve communication between practitioners and the public.

Analyses using academic spoken corpora

Spoken corpora have a prominent role in the applied linguistic analysis of academic texts. The MICASE (<https://quod.lib.umich.edu/m/micase/>) was created in order to increase our understanding of the nature of academic speech and to make comparisons with academic writing. The corpus covers a wide range of spoken academic genres such as lectures, tutorials, seminars and practical demonstrations. It consists of about 200 hours of academic speech amounting to about 1.5 million words. The project has, for example, shown that the interactive nature of academic speech is closely related to conversation. This is reflected in a much greater use of discourse markers in academic speech than in academic writing. Swales and Malczewski (2001) found that *okay* or *now* was frequent in marking a new episode in academic speech events (‘new episode flags’). Mauranen (2001) studied how discourse-reflexive elements such as *let us talk about* and *finally* have an important role in organising the discourse of academic speech. The corpus data were also used to test the hypothesis that reflexivity involves power and authority over the discourse. A corresponding British corpus

is the BASE Corpus consisting of lectures and seminars recorded at the Universities of Warwick and Reading (<http://www2.warwick.ac.uk/fac/soc/al/research/collections/base/>).

Analyses using spoken learner corpora

Speaking is a particularly challenging task for learners of a language. It is therefore important to find out what skills are difficult to acquire and why. The pedagogical aspect of spoken corpora is at the forefront in the compilation of learner corpora to discover what learners actually say. Learner corpora are collections of authentic texts produced by foreign/second language learners stored in electronic form. A well-known example is *the International Corpus of Learner English* of argumentative essays written by advanced learners of English representing different nationalities. It is now also possible to use spoken learner corpora as a resource for studies in second language acquisition.

The aim of the *Louvain International Database of Spoken English Interlanguage* (LINDSEI) project (www.uclouvain.be/en-cecl-lindsei.html) is to provide a spoken counterpart of ICLE. The corpus consists of over a million words of transcribed interviews (and a short storytelling activity) and represents 11 different mother tongues: Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish and Swedish. The learner corpus contains a ‘learner profile’ making it possible to correlate the learners’ spoken language use with age, gender, education and visits to an English-speaking country.

Learner corpora have been used to study the range and variety of discourse markers in learners’ spoken production (Fung and Carter 2007), as well as specific discourse markers (Müller 2005; Aijmer 2009, 2011). The method is generally comparative. Fung and Carter (2007) contrasted data representing secondary learners at a school in Hong Kong with comparable data from the pedagogical sub-corpus in the CANCODE Corpus. The results showed large differences in frequency between the non-native speakers and the CANCODE speakers who represented the British English norm. To begin with, discourse markers were generally less frequent in the student data. Moreover, the Hong Kong students did not use the same discourse markers as British speakers generally use. In particular, learners avoided discourse markers for interpersonal functions. Similar results have been shown in studies focusing on a single discourse marker. The differences between learners and native speakers with regard to forms, frequencies and functions can then be the basis for a more systematic teaching of pragmatic phenomena such as discourse markers.

Aijmer (2011) used the spoken component of the LINDSEI Corpus to study Swedish advanced learners’ use of *well* and compare it with a similar corpus of native English speakers. Swedish students were seen to underuse *well*, and they used it more for speech management purposes (e.g. when they were at a loss for words and needed time for planning). On the other hand, they used *well* less for interpersonal functions such as qualifying an opinion (*well I think*) or disagreeing.

The *Trinity Lancaster Corpus (TLC)* (http://cass.lancs.ac.uk/?page_id=1327) is a new, still growing corpus of learner English with obvious pedagogical applications. It contains semi-formal speech elicited in an institutional setting as part of the examination of spoken English conducted by the Trinity College, London. The corpus provides transcribed interactions elicited in different speaker tasks between candidates and examiners at different levels of language proficiency. The L2 data were collected from speakers in six different countries: China, Italy, Mexico, Sri Lanka, India and Spain. Gablasova et al. (2017) used the corpus to examine expressions of epistemic stance across speakers representing different

L1 backgrounds and across different speaker tasks in the examination. The examination consisted of a monologic task where the L2 speaker talked about a topic of his or her own and several interactive tasks: discussion (of the topic presented by the examiner), an interactive task where the speaker takes an active role by asking questions and conversation. The authors adopted a multi-step approach to identify the forms expressing epistemic stance. In a preliminary step, they identified a list of candidate forms that were used as search items. A concordance was identified for each form, and a sample of the concordance forms was manually coded to check that the forms were used in an epistemic way. The largest difference was found between the monologic and dialogic tasks, but there were also differences in the distribution of stance markers in the interactive tasks. In addition, there was evidence of inter-speaker variation. The study confirms the importance of using corpora of different types of L2 speech to get a better understanding of the pragmatic competence of the learners. The pedagogic implications are wide-ranging with regard to methods helping speakers to master pragmatic skills, material development and testing.

The neglect of discourse markers in foreign language teaching seems to be a reality. However, the findings from learner corpus research can help us to better understand the problems of L2 learners and broaden our views on what has to be taught. For example, the pragmatic uses of *well* to mark disagreement or to soften a suggestion should be given some room for practice in the classroom to help students avoid misunderstanding and to become more fluent. Teaching can start with activities raising awareness in which learners are made to notice particular features of language, such as discourse markers, to talk about what they are doing in the discourse and make cross-language comparisons (Fung and Carter 2007: 434).

Corpora of the use of English as a lingua franca

English is now in worldwide use as a *lingua franca*. The speakers are not learners of English and do not look to native speakers as their norm. English as a *lingua franca* (ELF) operates under special conditions since it happens outside normal teaching activities. Corpora are crucial to find out what linguistic features are characteristic. Well-known corpora of ELF are the *Helsinki ELFA Corpus* (*Corpus of English as Lingua Franca in Academic settings*) (www.helsinki.fi/englanti/elfa/elfacorporus.html) and the *Vienna VOICE Corpus* (*Vienna Oxford Corpus of International English*) (www.helsinki.fi/varieng/CoRD/corpora/VOICE/). The ELFA Corpus consists of 1 million words and comprises university lectures, PhD defences and conference discussions where the speakers come from different mother tongue backgrounds. The VOICE Corpus comprises naturally occurring face-to-face interactions in English, where the speakers represent approximately 50 different languages. These corpora are needed to study whether there is something special about ELF and justifies regarding it as a special variety (House 2009a: 142).

Scholars have used the ELF corpora to look at the way the interactants manage turn-taking, topic change and misunderstanding of utterances (House 2009b). Another topic which has been studied is discourse markers. As House (2009b) shows, on the basis of authentic corpus interactions, ELF speakers used *you know* to create coherence and to gain time for planning rather than with the interpersonal function most previous studies have ascribed to it.

Corpora of speech-like varieties

'Historical speech' and scripted speech in talk shows differ from ordinary face-to-face conversation since they were not produced in naturalistic environments. As speech-related

varieties, they can be investigated through specialised types of spoken corpora. The corpora open up the possibility to study varieties of spoken language which have so far been relatively unexplored. The corpus data are used to study how these varieties differ from authentic language.

Historical corpora of dialogue and drama

The object of study in spoken corpus linguistics is ‘speech in interaction here and now’. Given that audio recording only goes back about a hundred years, we cannot study how language was spoken in the past. Collecting a corpus of data that bears some similarity to spoken language has therefore been the only option for scholars whose interests are in historical sociolinguistics or pragmatics. Culpeper and Kytö (2010) collected a corpus of historical dialogue texts likely to represent the ‘spoken face-to-face interaction of the past’. Their *Corpus of English Dialogues 1560–1760 (CED)* [see Culpeper and Kytö (2010)] comprises speech-related genres such as drama comedy, witness depositions, trial proceedings and parliamentary proceedings which are argued to be more closely related to speech than to writing. The advantages of these genres are said to be that they are (1) (re)presentations of spoken online language, (2) they are (re)presentations of face-to-face and relatively interactive language, (3) there is some coverage of a wide spectrum of social groups and (4) they are available in reasonably large quantities, depending on genre and period (Culpeper and Kytö 2010: 16). The method was to focus on phenomena that are speech-like in some respect and then investigate them quantitatively and qualitatively. For example, Culpeper and Kytö examined ‘pragmatic noise’ such as *oh* and *ah* which ‘is there to create an illusion of spokenness and to assist in signalling meanings in interaction’ (Culpeper and Kytö 2010: 222).

The authors found that some linguistic features occurred more often in their speech-related text types than in non-speech-related text types (*o*, *oh*, *fie*, *ah*, *alas*). There were also some speech-related text types which turned out to be more speech-like with regard to specific linguistic features. No speech-related text type was speech-like on every count. However, the case for play texts was the strongest (Culpeper and Kytö 2010: 401).

Corpora of soap operas

It is also possible to collect large amounts of conversational data from television. This has materialised in the *Corpus of American Soap Operas*, which comprises 100 million words of conversation (<http://corpus.byu.edu/soap/>). The corpus consists of scripted dialogues representing several different soap operas and spans the period 2000–2012. The dialogues in soap operas are scripted, but contain a large number of high-frequency linguistic features characteristic of informal spoken language. Quaglio (2009) compared transcripts of the American situation comedy *Friends* and the American English conversation part of the Longman Grammar Corpus focusing on particular linguistic features such as vague language, the personal expression of emotion and informal language (including swearing). The results suggested that although soap operas are supposed to replicate natural conversation, there were differences due to the medium. However, soap operas can also be explored as an object of study in themselves which help us to better understand, for example, how humour and melodrama are created in the interaction between the characters.

Spoken corpus-based studies in pragmatics and discourse

The use of language is traditionally studied in pragmatics and discourse on the basis of interactional data. Conversation analysis, for example, focuses on how speakers take turns in the conversation and how grammatical patterns emerge in the evolving interaction. However, this approach is, above all, suitable for a refined, fine-grained analysis of the interaction in smaller datasets. Corpus linguistics, on the other hand, is particularly good at describing large quantities of data in their linguistic context. However, it is now believed that the two traditions can learn from each other and that spoken corpora can be a resource for understanding more about elements which have pragmatic and discourse functions.

This rapprochement between two different traditions is reflected in the focus on lexical elements and patterns which can be searched for in the corpus and further analysed. Traditionally grammars have been characterised by their bias towards written language. However, currently ‘big’ grammars can hardly be called comprehensive unless they also include a section on forms and structures characteristic of spoken language. The *Cambridge Grammar of English* is devoted to describing the proportions between speech and writing in a balanced way (Carter and McCarthy 2006) and covers a number of phrases and key words which are prominent in spoken English. Similarly, the quantitative and qualitative differences between speech and writing are a topic permeating the *Longman Grammar of Spoken and Written English*. According to Biber and his research team, ‘there is a compelling interest in using the resources of the LSWE Corpus of spoken language transcriptions to study what is characteristic of the grammar of conversation’ (Biber et al. 1999: 1038). On the quantitative side, the corpus can be used to compare frequencies of linguistic elements across selected registers of speech and writing. The differences can then be linked to contextual and functional factors. The scope of a grammar of conversation ‘is immense’ (cf Biber et al. 1999: 1039) and includes topics which have previously been neglected.

For example, it is now common to draw attention to pragmatic phenomena of a lexical character which have been neglected because they are infrequent in writing. Biber et al. (1999) look at a class of ‘insert words’ in spoken discourse which operate outside the clause structure and do not fit into the canonical structures provided by written grammar. Inserts can be illustrated by discourse markers such as *anyway, cos, fine, good, great, like, now, oh, okay, right, so, and well* (list from Carter and McCarthy 2006: 214). Discourse markers tend to occur initially in the utterance or turn and can be marked prosodically as a separate tone unit. In the light of what we know about the organisation of spoken interaction, their characteristic properties are no longer ‘exceptional’ or ‘*ad hoc*’ but can get a principled explanation in terms of functional and contextual factors. Thus, discourse markers have pragmatic or discourse functions associated with marking boundaries in the discourse, hedging, politeness and involvement. Other topics which need to be discussed in spoken grammar are vocatives and types of address forms (e.g. Leech 1999), interjections (Norrick 2015), tag questions (Tottie and Hoffman 2006), left and right dislocation (Aijmer 1989) and conversational routines (*thanks, sorry*; Aijmer 1996).

Spoken corpora have also opened our eyes to features such as interruptions, repetitions, false starts, pauses, hesitations and self-corrections in spoken discourse and highlighted the fact that these phenomena are normal in informal conversation where speakers are continually faced with the need to plan online. As a result, these dysfluency phenomena are now generally viewed in positive terms as ‘planners’ (Tottie 2016) or as speech management phenomena (Allwood et al. 1990). Hesitation markers such as *er* and *erm* have been shown to perform functions in the discourse associated with turn-taking and changing

topics. According to Rühlemann (2007: 161), *er/erm* ‘attends to needs arising from co-constructions and discourse-management. It is an important means for the organization of turn-taking both as a turn-bidder and as turn-holder and managing discourse by marking new, focal information’.

Spoken corpus-based studies have also gone beyond single words in order to focus on ‘lexical bundles’ (Biber and Conrad 1999) or recurrent phrases (Altenberg 1998). Studies of recurrent patterns using corpora reveal that much of our spoken output consists of patterns which do not represent complete structural units. Biber and Conrad (1999) found, for example, that only 15 percent of lexical bundles in the spoken component of the LSWE were complete in terms of syntactic structure. The most frequent four-word lexical bundles in the corpus had a first-person subject followed by a stative verb and ended with the beginning of a complement clause (*I don’t think I, I don’t know what, I thought I would*) (Biber and Conrad 1999: 187). Adolphs and Carter (2013) carried out a similar study of what they called ‘clusters’ using the 10-million spoken component of the BNC. They focused, however, on those automatically extracted strings which displayed ‘pragmatic integrity’ in the sense that they are pragmatic categories rather than syntactic or semantic ones (Adolphs and Carter 2013: 27).

Corpus-assisted discourse analysis

Discourse analysis (where the term discourse is used in the sense of talking about a particular topic) can also be performed with sociopolitical aims. The dominant school of research bringing together ideology and corpora is referred to as corpus-assisted discourse analysis (Partington et al. 2013). Baker (2006), in particular, has shown that the techniques of corpus investigation can be used to analyse the discourses about a topic and to reveal hidden messages which are not open to introspection. Discourses ‘are not valid descriptions of people’s “beliefs” or “opinions” and they cannot be taken as representing an inner, essential aspect of identity such as personality or attitude. Instead they are connected to practices and structures that are lived out in society from day to day’ (Baker 2006: 4). The corpus techniques involve the identification of recurrent lexical items, phrases and collocations in order to unpack the beliefs and assumptions which underlie what is actually said. The corpus is usually ‘specialised’ to make it possible to study a particular topic. Baker, for example, analysed different sides on the issue of fox hunting in debates in the House of Commons using a corpus built for this purpose. To begin with, he collected electronic transcripts of the debates which took place in 2002 and 2003. As a first step, he created lists of words which were frequently used either by the pro- or anti-hunt protagonists. The lists were used to produce specific key word lists exhibiting words which were more frequently used by the anti-hunt supporters than by those supporting the hunt. The key words were then further submitted to an analysis of concordance and collocational search, which showed some interesting patterns which could be identified with a pattern for or against fox hunting.

More recently, Taylor (2011) used the corpus-assisted discourse analysis approach to analyse the Hutton inquiry in the UK from the point of view of (im)politeness. The inquiry investigated the circumstances of the death of David Kelly, a leading scientific adviser on weapons of mass destruction. A corpus was created using the official transcriptions of the inquiry. Different examination types could be distinguished depending on whether they were friendly or hostile. WordSmith tools were used for the key word analysis of the corpus data, which were further analysed for (im)politeness.

The forms of analysis and the interest in how issues of gender, class and power are reflected in a particular text are also characteristic of critical discourse analysis (Fairclough

1995). However, critical discourse analysis (Fairclough 1995) has generally shied away from the use of corpora and does not formulate research questions so that they can be answered on the basis of corpora. Studies in corpus-assisted discourse studies (CADS), such as Baker's study, typically delve into sociopolitical issues through the use of corpora. CADS has, for example, studied how the conflict in Iraq was discussed and reported in the Senate and in the British Parliament using specialised corpora (Morley and Bayley 2009). The term corpus-assisted (rather than corpus-driven) refers to the fact that there are parts of the analysis which cannot be carried out by means of corpus techniques and methods but require a close reading of the texts. However, the questions asked about language and ideology are associated with a critical academic enquiry and an analysis of society.

Future directions

Spoken corpora and corpus linguistic methods are a good starting point for a discussion of what linguistics has to offer to digital humanities and the kind of research that can be carried out. Different types of spoken corpora have had a phenomenal growth in the last few decades and are also applied in areas of research seeking a better understanding of human society or concerned with a critical analysis of its institutions. A broad range of topics and corpora have been suggested in this chapter which are of common concern to digital humanities and corpus linguistics. The topics include language teaching, healthcare, 'historical' speech and culture, new spoken media, critical discourse analysis and language variation. We can expect continued progression in these and other areas also to be reflected in research undertaken under the umbrella of digital humanities. This closeness also puts a pressure on corpus linguists to make their corpus research more visible and to present the data in a user-friendly way.

Looking to the future, new developments in this area involve moving from monomodal spoken corpora into the multimodal domain. We are now witnessing a wave of interest in exploring corpus methods to capture, store and use non-verbal data for research. A pioneering enterprise is the building of the *Nottingham Multimodal Corpus (NMMC)* developed by a team of people already well-known for their corpus-based studies of spoken language (see e.g. Knight et al. 2008). The corpus was compiled between 2005 and 2008 and consists of 250,000 words recorded in academic settings. The project extends the scope of corpus linguistic techniques by also marking up and classifying gestures and other non-verbal features interacting with the verbal communication. Another part of the corpus design is the use of special software tools that can synchronise data with a particular encoding or transcription (Adolphs and Carter 2013: 151).

One way of understanding the potential of multimodal corpora for research is to consider the study carried out by some members of the research team relating head nods to listener responses (Adolphs and Carter 2013; Knight 2011b). The listener's responses, or 'backchannels', are exemplified by *yeah*, *right* and *I see*. Backchannels are inserted by the listener 'here and now' in the unfolding discourse monitoring ongoing talk. Depending on their functions in spoken communication, they can, for example, mark the point in the conversation where the listener has received adequate information (information receipt) or a more involved response (engaged response token). In order to describe how verbal and non-verbal features interact, it is also important to find a coding scheme for describing head nods. Head nods can be short or long, and they can differ with regard to intensity and placement (the points in the discourse where they occur). It may then be possible to assign a discourse function to a backchannel which can be linked to information about the type of head nod.

However, compared to monomodal corpora, the development of multimodal corpora is still in its infancy. Multimodal corpora need to be made bigger and made to represent a wider range of text types. Other items on the ‘wish list’ would be better methods and tools for recording and representing multiple forms of data, which would make it easier and quicker to compile a multimodal corpus. Moreover, software tools are needed that build on and extend tools existing for the analysis of monomodal corpora, such as WordSmith tools and Sketch Engine (Knight 2011a).

Further reading

- 1 McEnery, T. and Hardie, A. (2012). *Corpus linguistics*. Chapter 9. Conclusion. Cambridge: Cambridge University Press.
The authors present arguments (especially in 9.5) why disciplines in the humanities and the social sciences can benefit from corpora and corpus methods.
- 2 McEnery, T. and Hardie, A. (2012). *Corpus linguistics*. Chapter 2. Accessing and analysing corpus data.
The chapter gives an overview of mark-up conventions, methods, tools and statistical techniques in corpus linguistics.
- 3 Adolphs, S. and Carter, R. (2013). *Spoken corpus linguistics: From monomodal to multimedial*. Chapter 1. Making a start. Building and analysing a spoken corpus.
The chapter provides an overview of some of the main approaches to spoken corpus research. Among the topics discussed are the design of spoken corpora, issues involving the transcription and annotation of spoken data, the role of metadata accounting for the speakers and the types of activity.

References

- Adolphs, S., Brown, B., Carter, R., Crawford, P. and Sahota, O. (2004). Applying corpus linguistics in a health care context. *Journal of Applied Linguistics* 1(1): 9–28.
- Adolphs, S. and Carter, R. (2013). *Spoken corpus linguistics: From monomodal to multimodal*. New York: Routledge.
- Aijmer, K. (1989). Themes and tails: The discourse functions of dislocated elements. *Nordic Journal of Linguistics* 12(2): 137–154.
- Aijmer, K. (1996). *Conversational routines in English: Convention and creativity*. London: Longman.
- Aijmer, K. (2009). “So er I just sort of I dunno I think it’s just because . . .”: A corpus study of ‘I don’t know’ and ‘dunno’ in learner spoken English. In A.H. Jucker, D. Schreier and M. Hundt (eds.), *Corpora: pragmatics and discourse. Papers from the 29th international conference on English language*. Amsterdam and New York: Rodopi, pp. 151–168.
- Aijmer, K. (2011). Well I’m not sure, I think . . . The use of *well* by non-native speakers. *International Journal of Corpus Linguistics* 16(2): 231–254.
- Aijmer, K. (2015). Analysing discourse markers in spoken corpora: *Actually* as a case study. In P. Baker and T. McEnery (eds.), *Corpora and discourse studies. Integrating discourse and corpora*. London and New York: Palgrave Macmillan, pp. 88–109.
- Aijmer, K. (2018). ‘*That’s well bad*’. Some new intensifiers in Spoken British English. In V. Brezina, R. Love and K. Aijmer (eds.), *Corpus approaches to contemporary British speech*. London and New York: Routledge, pp. 60–95.
- Allwood, J., Nivre, J. and Ahlsén, E. (1990). Speech management – On the non-written life of speech. *Nordic Journal of Linguistics* 13: 3–48.
- Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A.P. Cowie (ed.), *Phraseology. Theory, analysis, and applications*. Oxford: Oxford University Press, pp. 101–122.
- Andersen, G. (2001). *Pragmatic markers and sociolinguistic variation*. Amsterdam and Philadelphia: John Benjamins.

- Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.
- Beeching, K. (2016). *Pragmatic markers in British: Meaning in social interaction*. Cambridge: Cambridge University Press.
- Biber, D. and Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgård and S. Oksefjell (eds.), *Out of corpora: Studies in honour of Stig Johansson*. Amsterdam: Rodopi, pp. 181–189.
- Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999). *The Longman grammar of spoken and written English*. London: Longman.
- Carter, R. and McCarthy, M. (2006). *The Cambridge grammar of English*. Cambridge: Cambridge University Press.
- Cheng, W., Greaves, C. and Warren, M. (2008). *A corpus-driven study of discourse intonation. The Hong Kong Corpus of Spoken English (prosodic)*. Amsterdam and Philadelphia: John Benjamins.
- Culpeper, J. and Kytö, M. (2010). *Early modern English dialogues: Spoken interaction as writing*. Cambridge: Cambridge University Press.
- Davies, M. (2010). The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25(4): 447–464.
- Erman, B. (1992). Female and male usage of pragmatic expressions in same-gender and mixed-gender interaction. *Language Variation and Change* 4(2): 217–234.
- Fairclough, N. (1995). *Critical discourse analysis*. London: Longman.
- Fung, L. and Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics* 28(3): 410–439.
- Gablasova, D., Brezina, V., McEnery, T. and Boyd, E. (2017). Epistemic stance in Spoken L2 English: The effect of task and speaker style. *Applied Linguistics* 38(5): 613–637.
- Gabrielatos, C., Torgersen, E., Hoffmann, S. and Fox, S. (2010). A corpus-based sociolinguistic study of indefinite article forms in London English. *Journal of English Linguistics* 38(4): 297–334.
- House, J. (2009a). Introduction: The pragmatics of English as a Lingua Franca. In J. House (ed.), *English Lingua Franca. Special issue of Intercultural Pragmatics* 6(2): 141–145.
- House, J. (2009b). Subjectivity in English as lingua franca discourse. The case of *you know*. In J. House (ed.), *English Lingua Franca. Special issue of Intercultural Pragmatics* 6(2): 171–193.
- Kallen, J.L. and Kirk, J.M. (2012). *SPICE-Ireland: A user's guide*. Belfast: Cló Ollscoil na Banríona.
- Kirk, J.M. (2016). The pragmatic annotation scheme of the SPICE-Ireland Corpus. *International Journal of Corpus Linguistics* 21(3): 299–323.
- Kirk, J.M. and Andersen, G. (2016). Compilation, transcription, markup and annotation of spoken corpora. *International Journal of Corpus Linguistics* 3: 291–298.
- Knight, D. (2011a). The future of multimodal corpora: O futuros dos corpora modais. *Revista Brasileira de Linguística Aplicada* 11(2): 391–415.
- Knight, D. (2011b). *Multimodality and active listenership: A corpus approach*. London: Bloomsbury.
- Knight, D., Adolphs, S., Tennent, P. and Carter, R. (2008). *The Nottingham multi-modal corpus. A demonstration*. Paper delivered as part of the ‘Multimodal Corpora’ workshop held at the 6th Language Resources and Evaluation Conference, Palais des Congrès Mansour Eddahbi, Marrakech, May, Morocco.
- Kretzschmar, W.A., Anderson, J., Beal, J.C., Corrigan, K.P., Opas-Hänninen, L.L. and Plichta, B. (2006). Collaboration on corpora for regional and social analysis. *Journal of English Linguistics* 34: 172–205.
- Knowles, G. (ed.) (1996). *A corpus of formal British English speech: The Lancaster/IBM spoken English corpus*. London: Longman.
- Labov, W. (1966). The linguistic variable as structural unit. *The Washington Linguistics Review* 2: 4–22.
- Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: Philadelphia University Press.
- Leech, G. (1999). The distribution and function of vocatives in American and British English conversation. In H. Hasselgård and S. Oksefjell (eds.), *Out of corpora: Studies in honour of Stig Johansson*. Amsterdam: Rodopi, pp. 107–118.

- Lu, X. (2010). What can corpus software reveal about language development? In A. O’Keeffe and M. McCarthy (eds.), *The Routledge handbook of corpus linguistics*. London and New York: Routledge, pp. 184–193.
- Mauranen, A. (2001). Reflexive academic talk: Observations from MICASE. In R. Simpson and J. Swales (eds.), *Corpus linguistics in North America*. Ann Arbor: University of Michigan Press, pp. 165–178.
- McCarthy, M. (1998). *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.
- McCarthy, M. and Carter, R. (2001). Size isn’t everything: Spoken English, corpus, and the classroom. *Tesol Quarterly* 35(2): 337–340.
- McEnery, T. and Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
- Morley, J. and Bayley, P. (eds.) (2009). *Corpus-assisted discourse in the Iraq conflict: Wording the war*. London: Routledge.
- Müller, S. (2005). *Discourse markers in native and non-native English discourse*. Amsterdam and Philadelphia: John Benjamins.
- Norrick, N. (2015). Interjections. In K. Aijmer and C. Rühlemann (eds.), *Corpus pragmatics: A handbook*. Cambridge: Cambridge University Press, pp. 249–275.
- Partington, A., Duguid, A. and Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS)*. Amsterdam and Philadelphia: John Benjamins.
- Quaglio, P. (2009). *Television dialogue: The sitcom Friends vs. natural conversation*. Amsterdam and Philadelphia: John Benjamins.
- Rühlemann, C. (2007). *Conversation in context: A corpus-driven approach*. London and New York: Continuum.
- Rühlemann, C. and Aijmer, K. (2015). Corpus pragmatics: Laying the foundations. In K. Aijmer and C. Rühlemann (eds.), *Corpus pragmatics: A handbook*. Cambridge: Cambridge University Press.
- Rühlemann, C. and O’Donnell, M.B. (2012). Introducing a corpus of conversational narratives: Construction and annotation of the narrative corpus. *Corpus Linguistics and Linguistic Theory* 8(2): 313–350.
- Simpson, J. (2010). Applied linguistics in the contemporary world. In J. Simpson (ed.), *Routledge handbook of applied linguistics*. London and New York: Routledge, pp. 1–7.
- Stenström, A-B. (2006). The Spanish discourse markers *o sea* and *pues* and their English correspondences. In K. Aijmer and A-M. Simon-Vandenberg (eds.), *Pragmatic markers in contrast*. Amsterdam: Elsevier, pp. 155–172.
- Stenström, A-B., Andersen, G. and Hasund, I.K. (2002). *Trends in teenage talk*. Amsterdam and Philadelphia: John Benjamins.
- Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*. Cambridge, MA: Blackwell.
- Svartvik, J. (1990). *The London-Lund corpus of spoken English. Description and research*. Lund: Lund University Press.
- Swales, J. and Malczewski, B. (2001). Discourse management and new-Episode flags in MICASE. In R. Simpson and J. Swales (eds.), *Corpus linguistics in North America*. Ann Arbor: University of Michigan Press, pp. 145–164.
- Tagliamonte, S. (2016). *Teen talk: The language of adolescents*. Cambridge: Cambridge University Press.
- Taylor, C. (2011). Negative politeness features and impoliteness: A corpus-assisted approach. In B. Davies, A. Morrison and M. Haugh (eds.), London: Continuum, pp. 209–231.
- Terras, M. (2011). Present, not voting: Digital humanities in the panopticon. Closing plenary speech, digital humanities 2010. *Literary and Linguistic Computing* 26(3): 267–269.
- Terras, M. (2016). A decade in the digital humanities. *Journal of Siberian Federal University. Humanities & Social Sciences* 7: 1637–1650.

- Torgersen, E., Gabrielatos, C., Hoffmann, S. and Fox, S. (2011). A corpus-based study of pragmatic markers in London English. *Corpus Linguistics and Linguistic Theory* 7(1): 93–118.
- Tottie, G. (2016). Planning what to say: *Uh* and *um* among the pragmatic markers. In G. Kaltenböck, E. Keizer and A. Lohmann (eds.), *Outside the clause*. Amsterdam and Philadelphia: John Benjamins, pp. 97–122.
- Tottie, G. and Hoffmann, S. (2006). Tag questions in British and American English. *Journal of English Linguistics* 34(4): 283–311.
- Wallis, S.A. (2014). What might a corpus of parsed spoken data tell us about language? In L. Veselovská and M. Janebová (eds.), *Complex visibles out there: Proceedings of the Olomouc linguistics colloquium 2014: Language use and linguistic structure*. Olomouc: Palacký University, pp. 641–662. pre-published (PDF).
- Weisser, M. (2015). Speech act annotation. In K. Aijmer and C. Rühlemann (eds.), *Corpus pragmatics. A handbook*. Cambridge: Cambridge University Press, pp. 84–110.
- White, P.R., Simon-Vandenberg, A.-M. and Aijmer, K. (2007). Presupposition and ‘taking-for-granted’ in mass communicated political argument: An illustration from British, Flemish and Swedish political colloquy. In A. Fetzer and G. Lauerbach (eds.), *Political Discourse in the Media*. John Benjamins: Amsterdam and Philadelphia, pp. 31–74.
- Wichmann, A. (2004). The intonation of *please*-requests: A corpus-based study. *Journal of Pragmatics* 36: 1521–1549.

Written corpora

Sheena Gardner and Emma Moreton

Introduction

The humanities has always been concerned with written texts, but the digital revolution has allowed us to electronically store, annotate and analyse ever-increasing amounts of data in new and creative ways. Indeed, the digital humanities (using digital tools and techniques, including corpus linguistic methods to address humanities-related research questions) offers an opportunity to bring together scholars from across the disciplines to look at ways of harnessing the power of new technologies in the study of humanities data. Here we focus on corpora of written texts.

Collections of written texts, or archives, which are simply assembled and may or may not be indexed and catalogued are differentiated from corpora, which involve a careful sampling of texts that are representative, for instance, of written English and are digitally prepared for analysis. Corpora are generally understood to be ‘collections of naturally-occurring language data, stored in electronic form, designed to be representative of particular types of text and analysed with the aid of computer software tools’ (Nesi 2016: 206). Corpora include not only the actual texts but also metadata, or information about the production of the texts (e.g. who, where, when, why, about what), which is used to associate the text with its social context and to assist in comparative analyses within and across corpora.

Written corpora can be analysed from a range of quantitative perspectives to uncover patterns that would be hard to detect manually. These might be based on frequency, co-occurrence or relative frequency. For example, frequency data might be sought to answer questions such as ‘Which nations do Shakespeare and Dickens refer to most in their writing?’ Or co-occurrence queries might be used for more complex questions, such as ‘Which images are typically associated with darkness in Shakespearean vs. Dickensian writing?’ This latter question might be answered quantitatively in terms of most frequent collocating items, which in turn might simply reflect the nature of English. For example, we can identify which adjectives most frequently occur with a noun such as ‘night’ and compare the co-location findings for Shakespeare and Dickens. Collocation search parameters can be set to count only those that occur immediately before the noun or to include those that occur up to a certain number of words before and/or after the noun. Alternatively, the question could

be answered in relative terms by comparing items that are distinctive, or ‘key’, in the writing of each author when compared with a general English corpus. These central concepts of frequency, collocation and keyness are widely used and explained in all introductions to corpus linguistics (e.g. Hunston 2002). More complex ‘query language’ would be needed to answer questions such as ‘How are women represented in Shakespeare and Dickens?’, and as the results are unlikely to be as directly comparable, an interpretation of the results will normally involve grouping the results into categories, which may be linguistic (e.g. semantic categories or categories of metaphor) or may derive from theories of gender representation or literary criticism.

Research on written corpora involves using tools to access and manipulate the information in corpora in ways that can answer questions, particularly those related to frequency, phraseology and collocation within and across texts, that our intuitions about language or our linear reading of texts cannot. Around this quantitative methodological core, and the more qualitative and interpretive examination of significant or key items (words, phrases or collocations) and the contexts in which they occur, the questions posed and how findings are explained are as varied as research in humanities itself.

Our aim in this chapter is to describe influential written corpora in English. We present a broad categorisation (Table 3.1), followed by an account of historical developments in the field and issues related to corpus contents and the growing statistical sophistication of analytical tools. We then present two small-scale studies: a corpus search for the term *humanities* in several large, ready-made, freely available corpora and an investigation of features of correspondence in three specialist corpora. The chapter concludes with suggestions for future directions, further reading, related topics and references. Our aim is to demonstrate how written corpora, and corpus methods of analysis, complement and contribute to the multidisciplinary field of digital humanities.

Types of corpus

The categorisation in Table 3.1 is intended to give a broad overview of types of written corpus that exist in English. Similar classifications are found in most introductions to corpus linguistics (e.g. Huber and Mukherjee 2013; Hunston 2002: 14–16; the Lancaster University corpus linguistics MOOC session 2), as well as on websites that offer access to multiple corpora, such as Brigham Young University (<http://corpus.byu.edu>) and Sketch Engine (www.sketchengine.co.uk/user-guide/user-manual/corpora/corpus-types/).

This classification is not exhaustive; other categories include monolingual (most of the corpora listed are monolingual English), multimodal (e.g. transcripts plus recordings), reference (large general corpora are often used as reference corpora against which smaller bespoke corpora can be compared), pedagogic (e.g. all the receptive texts for a learner – textbooks, lectures, etc.) and regional (the International Corpus of English [ICE] was designed to compare texts from 20 varieties of English, while the SCOTS corpus focuses on a single variety). Nevertheless, it provides an orientation to the scope of written corpora in English.

The earliest national corpora are the Lancaster-Oslo-Bergen (LOB) corpus of British English and the Brown corpus of American English, both developed in projects that began in the 1960s. These were succeeded by the Bank of English in the 1980s (see information on COBUILD later) and other larger national corpora, specifically the British National Corpus (BNC) developed between 1991 and 1994 and the Corpus of Contemporary American (COCA), both of which are widely used today. The BNC is valued for its extensive annotation and multiple categories. It reflects an adult educated variety of English with

Table 3.1 Types of corpus exemplified

Type	Contents	Examples
General	A variety of texts intended to reflect general uses of spoken and written language	<i>The British National Corpus</i> (BNC), a general corpus of British English <i>The Corpus of Contemporary American English</i> (COCA), a general corpus of American English
Historical/ Diachronic	From specific periods, or allows comparison by, e.g. decade	<i>Hansard corpus</i> (1803–2005) <i>Time Magazine corpus</i> (1923–2006) <i>The Siena Bologna/Portsmouth Modern Diachronic Corpus</i> (SiBol/Port) contains 787,000 articles (385 million tokens) from three UK broadsheet newspapers in 1993, 2005 and 2010
Learner	Texts produced by learners	<i>International Corpus of Learner English</i> (ICLE) with 3 million words of argumentative essays <i>Cambridge Learner Corpus</i> (CLC) with 50 million words from Cambridge exam scripts submitted by 220,000 students from 173 countries
Monitor	Regularly updated to track changes in a language	<i>The Bank of English</i> (BoE) (see discussion at http://corpus.byu.edu/coca/compare-boe.asp) <i>News on the Web</i> (NOW) corpus dates from 2010 and grows by about 10,000 web articles each day
Parallel	Texts and their aligned translation(s)	<i>EUR-LEX</i> is a multilingual parallel corpus of European Union documents translated into the official European languages
Multilingual	Containing texts in several languages	<i>European Corpus Initiative (ECI) Multilingual Corpus</i> , containing texts (mainly newspapers and fiction) in over 20 languages
Specialist	Designed for a specific purpose	<i>British Academic Written English</i> (BAWE) contains 6.5 million words of assessed university student writing across disciplines and levels of study <i>Hong Kong Professional Corpora</i> specialises in financial services, engineering, governance, etc. See also ‘Small correspondence corpora’ in the text
Web Corpus	Created by web crawling, categorised by web domain, typically the most synchronous of corpora	<i>ententen</i> (2013) contains 19 billion words of English on the web <i>GloWbE</i> contains 1.9 billion words from websites in 20 different English-speaking countries

90 percent written texts, including media, literary and academic sources across domains, and 10 percent transcripts of spoken language. Broadly speaking there are massive corpora compiled by web crawling. For example, Sketch Engine houses TenTen corpora that aim to be more than 10 billion words each for more than 30 languages and a news corpus of 28 billion words that grows by 800 million words a month (www.SketchEngine.eu/timestamped-english-corpus, accessed 17/04/2018). There are large (COCA had over 560 million words in December 2017) general corpora, such as the BNC and COCA, that collect a range of genres according to a specific design; there are corpora built on existing archives or collections, such as the EUR-LEX, Hansard, *Time Magazine* and CLC corpora; and corpora that are compiled for specific research purposes, such as the ICLE and British Academic Written

English (BAWE) corpora. Corpora also vary in the nature and extent of their annotation or mark-up, and thus how amenable they are to answering specific questions in the humanities. Our two sample analyses later provide some insights into this, as do other chapters in the handbook.

Background

Research across the humanities has long been informed by the analysis of written texts. Pre-digital studies used techniques that are familiar today. Concordances of texts such as the complete works of Shakespeare and the King James Bible were created manually (see Wisby 1962) by searching for patterns and extracting items with similar functions, references or meanings and published as reference books.

In linguistics, Peter Fries (2010) describes Charles Fries's systematic collection, scientific analysis and counting of relative frequencies of paradigmatic features (e.g. intonation patterns associated with yes–no questions) in 'a representative sample of the language of some speech community'. Fries Junior argues that, unlike the transformational linguistics that came to dominate the field particularly in America, the early twentieth-century linguistics of Firth, Quirk, Fries and others was a natural precursor to current corpus linguistics. This explains Thompson and Hunston's identification of three concerns shared by systemic-functional linguistics and corpus linguistics: a focus on 'naturally occurring language, the importance of context (or contextual variation), and of frequency' (2006: 5).

Similarly in folklore and literature studies, Michael Preston (n.d.) describes the development of computerised concordances:

Computer-assisted study of folklore and literature was initiated shortly after World War II by Roberto Busa, S.J., who began preparing a concordance to the works of Thomas Aquinas in 1948, and Bertrand Bronson, who made use of the technology to study the traditional tunes of the Child ballads. In the pre-computer era, Bronson had worked with punched cards which he manipulated with a mechanical sorter and a card-printer. Many early efforts at producing concordances by computer were modeled on the punched card/sorter/reader-printer process, itself an attempt at mechanizing the writing of slips by hand which were then manually sorted.

What have changed are the speed and ease with which we can now collate large quantities of written texts and, when they are digitally transcribed and marked up, the speed and ease with which we can analyse them.

Two of the earliest digital corpus projects were the Brown Corpus of American English (www.helsinki.fi/varieng/CoRD/corpora/BROWN/ accessed 17/04/2018) and its counterpart, the LOB corpus of British English. Developed in the 1960s, these each contained 500 texts of around 2000 words distributed across 15 text categories. This inspired the compilation of further corpora with the same design that enabled ready comparison across varieties of English. Such developments whetted researchers' appetites to develop bigger and more diverse corpora.

An ambitious project was conducted at the University of Birmingham under the leadership of John Sinclair. The project involved not only the construction of a very large corpus of contemporary English – it is currently around 450 million words (www.titania.bham.ac.uk/docs/svenguide.html) – but also the development of methods and approaches

to corpus analysis that changed the way we work with English text and still pervade the field today. The annual Sinclair Lectures regularly remind us not only of the achievements of the COBUILD project (see later), and the related advances in lexicography in particular, but also that when it started in the 1980s, it was the English Department, and his project, that had the largest and most expensive computer on campus. The project was funded by Collins and created to inform the making of dictionaries, grammar books and related teaching materials.

In the last 50 years, we have moved from mainframe computers, with data entered on file cards, reminiscent of the handwritten cards used by Charles Fries. In the 1980s these were replaced by desktop computers, which allowed individual researchers to develop and analyse their own digital corpora using concordancing software, and today many hundreds of digital corpora of written texts are increasingly available for online use with sophisticated and user-friendly software programmes by students from mobile devices.

Critical issues and topics

Much humanities research involves analysing and interpreting texts, and it is therefore not surprising that researchers are attracted to digital analyses of written corpora to seek evidence in or interpretation of texts to support their theories. There are many decisions to be made in compiling a corpus, however, and when a corpus has been designed for one purpose, it may not be appropriate for use in different contexts. As the number of publicly available corpora increases, the methods used to compile them come under increasing scrutiny. When the COBUILD project began, the aim was to collect large numbers of words, and so novels and academic journals were an easy source of written data. Equally, when an academic corpus was being developed by Averil Coxhead in order to identify items for an academic word list (AWL) (www.victoria.ac.nz/lals/resources/academicwordlist/information/corpus), corpus collection was heavily influenced by the textbooks at one university and is biased more towards business than sciences, for instance. This and other critiques have led to alternative corpora to explore in the production of new AWLs (e.g. see review in Liu 2012). Nevertheless, both the COBUILD and the AWL projects have significantly influenced English language teaching and have inspired the development of more ‘representative’ corpora.

There are, of course, issues of what is ‘representative’, and perhaps a more realistic goal for many is to develop a ‘balanced’ corpus. By identifying particular features of interest in advance (e.g. male vs. female, humanities vs. sciences, quantitative vs. qualitative research, correspondence to subordinates, to superiors, or to peers), it should be possible to balance the corpus in these respects. It will never be possible to have corpora that are balanced for all situational and contextual variables. So, corpus developers should be clear about their research goals when they are developing the corpus, and corpus users should seek to understand not only the composition of the corpus but also the rationale for its construction, which features are ‘balanced’, and in what ways.

A third critical issue relates to the statistical sophistication needed to interrogate and use corpus findings. The statistics exemplified in areas such as multidimensional analysis (Biber 1988) and the use of R (Gries 2013) is crucial to the development of corpus linguistics as a field, but these are not traditional strengths of researchers in humanities, and the extent to which those investigating corpus resources need to understand, or teach, the statistics behind the findings is a moot point.

Current contributions and research

Corpora are used in many areas of English language research. They inform theories of language and of language learning and teaching (e.g. Frankenberg-Garcia et al. 2011), translation, second language acquisition and (critical) discourse analysis. For instance, Caldas-Coulthard and Moon (2010) use corpora to help ‘deconstruct hidden meanings and the asymmetrical ways people are represented in the press’ (2010: 99).

Here we illustrate the scope of written corpora and corpus research through two key domains: diachronic English linguistics and English language teaching materials.

Diachronic/historical corpora

In terms of historical written corpora, a notable starting point is *A Representative Corpus of Historical English Registers* (ARCHER). Initially constructed in the 1990s by Biber and Finegan (Biber et al. 1994), ARCHER is a ‘multi-genre historical corpus of British and American English covering the period 1600–1999’. Other multi-genre diachronic corpora include the *Corpus of Historical American English* (COHA), which includes over 400 million words dating from the 1810s–2000s, and the *Corpus of English Religious Prose* containing a range of genres (including prayers, sermons, treatises and religious biographies), dating from 1150 to the end of the eighteenth century. Genre-specific corpora include the recently released *Corpus of Historical English Law Reports 1535–1999* (CHELAR) – 185 files (369 texts) totalling 463,009 words – and the *Zurich English Newspaper Corpus* (ZEN), consisting of 349 complete newspaper issues containing 1.6 million words.

Arguably, some of the most useful materials for exploring language change and variation are ego documents: diaries, first-person narratives and private correspondence. The *Corpus of Early English Correspondence* (CEEC), produced at the University of Helsinki, comprises 188 letter collections (12,000 letters) dating from c. 1403 to 1800. The resource serves as an excellent example of how sociobiographic and extra-linguistic metadata might be captured in a systematised way, allowing users to explore language in relation to variables such as the author’s sex, rank, education and religion (Raumolin-Brunberg and Nevalainen 2007: 162). Researchers at VARIENG (the Research Unit for the Study of Variation, Contacts and Change in English, at the University of Helsinki) have used this material to explore, for instance, expressions of politeness, spelling variation and lexical change.

Additionally, a growing number of digital humanities projects are now using letters as a primary data source. With some basic pre-processing work, these digitised correspondence collections can be used for linguistic analysis. For example, the *Mapping the Republic of Letters* project at Stanford University, in collaboration with various partners including Oxford University’s *Cultures of Knowledge* project, maps networks of correspondence between ‘scientific academies’ within Europe and America during the seventeenth and eighteenth centuries to explore how such networks facilitated, amongst other things, the development and dissemination of ideas and the spread of political news.

Other projects include the *Electronic Enlightenment* project, at the University of Oxford, containing over 69,000 letters and documents from the early modern period and the *Darwin Correspondence Project* at Cambridge University, which has collected and digitised roughly 15,000 letters by Charles Darwin, providing information ‘not only about his own intellectual development and social network, but about Victorian science and society in general’. Finally, the *Victorian Lives and Letters Consortium*, coordinated by the University

of South Carolina, has brought to light samples of life-writing from the period spanning the coronation of Queen Victoria to the outbreak of World War I. Documents include letters by Thomas Carlyle and the diaries of John Ruskin.¹

Smaller-scale projects include Sairio's (2009) study of correspondence by Elizabeth Montagu (as part of her research on letters of the bluestocking network) and Tieken-Boon van Ostade's (2010) study of unpublished correspondence of Robert Lowth. Other digital collections include *The Browning Letters* – 574 letters between Victorian poets Robert Browning and Elizabeth Barrett Browning; *Bess of Hardwick's Letters* – 234 letters dating from c. 1550 to 1608 all available in Extended Markup Language (XML) format; and the *Van Gogh Letters Project* – around 800 letters by van Gogh to his brother Theo as well as to artist friends, including Paul Gauguin and Emile Bernard (translated into English).²

Personal and corporate letters of eminent persons (e.g. public political and social figures) have long been used for social, historical and cultural studies. Such letters are saved, transcribed, edited and published. Eminent persons themselves are sometimes part of this process – making copies of their own letters before sending them or retrieving letters from recipients. The practices of Thomas Jefferson, for example, have received much attention: his efforts to repair the research record and his use of the letterpress and polygraph letter duplication systems (see Sifton 1977). Over the past decades, however, there has been a growing interest in what scholars have termed 'history from below' or 'intrahistoria'³ – that is, a history of the popular classes – ordinary men and women who experienced historical events and their consequences firsthand (see, for e.g. Amador-Moreno et al. 2016; Auer and Fairman 2012; Fairman 2012, 2015). For example, the *Letters of 1916* project is creating a crowdsourced digital collection of letters from the time of the Easter Rising (1915–1931) and the *Irish Emigration Database* (IED) housed at the Mellon Centre for Migration Studies in Northern Ireland, which contains over 4,000 letters by Irish migrants and their families dating from the seventeenth to the twentieth century and has been used to explore topics and themes in the discourse, as well as linguistic variation and change in Irish English.

Historical literary corpora are also being used to investigate language change and variation. The *Salamanca Corpus* contains literary texts from the 1500s to 1900s searchable by period, county, genre and author, allowing users to explore vernacular literature and the representation of dialect; the *York-Helsinki Parsed Corpus of Old English Poetry* contains 71,490 words of Old English text that has been parsed syntactically and morphologically, enabling linguistic analyses of various kinds; and the *Visualizing English Print* project (a collaboration between the University of Wisconsin-Madison Libraries, the University of Strathclyde and the Folger Shakespeare Library) has recently released plain-text files of early modern text and drama corpora. The variant detector VARD, a pre-processing tool designed to deal with spelling variations in historical data, has been used to standardise the texts for the corpus analysis work. VARD is also useful in other areas of humanities research (error tagging in learner corpora (Rayson and Baron 2011), the normalisation of spelling variations in text messages (Tagg et al. 2012) and the exploration of gender and spelling differences in Twitter and Short Message Service (SMS) (Baron et al. 2011)).

Other literary corpora include the *Corpus of English Novels* (CEN) – texts by 25 British and American authors dating from 1881 to 1922, and the *Shakespeare Corpus*, produced by Mike Scott, containing 37 plays, including separate files for all speeches by all characters within those plays. Finally, the *Charles Dickens Corpus* contains 14 texts (over 3 million words) and forms the basis of a collaborative project – *CLiC Dickens* – between the University of Nottingham and the University of Birmingham, which 'demonstrates through corpus

stylistics how computer-assisted methods can be used to study literary texts and lead to new insights into how readers perceive fictional characters' (*CLiC Dickens*).

English language teaching materials

A major strand in corpus research involves corporate collaboration to inform English language dictionaries, grammar books and teaching materials. The Collins Birmingham University International Language Database (COBUILD) project (Moon 2007) started by John Sinclair in the 1980s and informed by the Bank of English corpus, produced the COBUILD dictionary in 1987, followed by pattern grammars and teaching materials. Here we see corpus-driven research – corpora being used not to find current examples of grammatical and lexical usage, but to inspire a step-change in lexicography, a novel way of viewing language and a new theory of grammar as patterns. The COBUILD team produced teaching resources that challenged traditional notions of what to teach and how to teach it. For instance, compared to Hornby's (1974) 25 formal verb patterns, Francis, Hunston and Manning's pattern grammar (1996) identifies over 700 meaningful verb patterns; Willis (1994) demonstrates alternative explanations of three common grammatical features – the passive, the second conditional and reported statements – and reveals the value of lexical phrases. This interest in lexical phrases (or clusters, or n-grams, or bundles) has become a cornerstone of corpus linguistic research and has inspired new theories about how we learn language (e.g. Hoey's lexical priming, 2005).

The *Longman Grammar of Spoken and Written English* (Biber et al. 1999) builds on the Quirk et al. (1985) grammars and provides a descriptive grammar of English informed by the 40-million-word Longman Spoken and Written English (LSWE) corpus. One of its distinguishing features is the inclusion of register frequency information in spoken, media, fiction and academic texts. For example, although present-tense verbs are more common than past-tense verbs and are frequent in conversation and academic prose (for different reasons), in fiction past-tense verbs are more frequent (p. 456). Furthermore, some verbs, including *bet*, *know*, *mean* and *matter*, occur more than 80 percent in the present tense, while others, including *exclaim*, *pause*, *grin* and *sigh*, occur more than 80 percent in the past tense (p. 459).

A subsequent project was supported by the Educational Testing Service (ETS) to investigate the spoken and written registers that university students in the United States listen to or read, including classroom teaching, textbooks and service encounters (Biber et al. 2002). The data in the 2.7-million-word TOEFL 2000 Spoken and Written Academic Language (T2k-SWAL) corpus was collected from four sites across the United States and stratified by level and area of study. This design is well suited to a focus on the language needed to study at university, as assessed in TOEFL.

As these are essentially commercial projects, access to these corpora tends to be restricted, but in some cases it is possible to gain access on request.

Nowadays, many publishers employ their own corpus linguists to work on in-house corpora. Notable examples include Cambridge University Press, Longman, Oxford University Press and Pearson, all of which can use the works they publish in corpora developed for different purposes. For instance, the *Oxford Learner's Dictionary of Academic English* (Lea 2014) is informed by and includes extracts from the 85-million-word Oxford Corpus of Academic English (OCAE) which includes undergraduate textbooks, scholarly articles and monographs. The freely available Pearson academic collocates list was compiled from the written component of the 25-million-word Pearson International Corpus of Academic English (PICA), with the support of many well-known corpus linguists.

Other influential academic English corpora include those developed by Ken Hyland and used to develop a theory of metalanguage (Hyland 2005) and to provide evidence that supports arguments in favour of the disciplinary distinctiveness of academic English (and therefore for teaching English for specific academic purposes). Hyland (2008), for example, compares lexical bundles across disciplines in a 3.5-million-word corpus of research articles, master's and PhD theses from Hong Kong.

Freely available academic corpora include the 6.5-million-word ESRC-funded BAWE corpus of successful university student writing (www.coventry.ac.uk/BAWE). It has not only informed a genre classification (Nesi and Gardner 2012) and studies of academic English, including the use of shell nouns (Nesi and Moreton 2012), reporting verbs (Holmes and Nesi 2009), nominal groups (Staples et al. 2016), section headings (Gardner and Holmes 2009), Chinese writers (Leedham 2015) and lexical bundles (Durrant 2017), but it also underpins the Writing for a Purpose materials on the British Council LearnEnglish website (<https://learnenglish.britishcouncil.org/en/writing-purpose/>).

In this overview of current contributions and research, we have included just some of the written corpora, most of which are accessible online. More corpora are listed in the Further Reading databases and interfaces at the end of this chapter.

Sample analyses

Two contrasting sample analyses of written corpora are now presented.

Humanities in pre-loaded general, academic and historical corpora

Here we illustrate the nature of information that can be freely obtained from pre-loaded corpora on the open Sketch Engine and Brigham Young University (BYU) sites through searches for *humanities* in several popular corpora of written texts. The choice of *humanities* is not only appropriate in relation to the contents of this volume but also appropriate in that we consider the term to be a semantic shifter (Fludernik 1991); a term that has forms (signifiers) and referents, but whose meaning (signified) is understood according to the contexts in which it is used. The aim is not to suggest that this is how a study should be conducted, but to show through example the kinds of information that can be obtained from different corpora.

BNC

We started with a WordSketch search for the lemma *humanity* on the BNC at www.SketchEngine.co.uk/open. The results show that the only context in which it is frequent is as a postmodifier with 'of', where it occurred 432 times (3.85 pmw), compared to the next most frequent context 'in humanity' with 87 (0.77 pmw). Examples with 'of' include 'a sense of common humanity', 'the dregs of humanity', 'the whole of humanity' and 'in the interests of humanity'. As in these examples, it is widely (396 times, or 3.52 pmw) used as a singular, uncountable, abstract noun, 'humanity', rather than a plural form. Among collocates of the plural form 'humanities' is a group of educational terms (*humanities scholars, student, research, department, course*), as in this example, which is worth reading for the ideas it promotes as well as the use of the term 'humanities'.

What will the next decades hold? Will it be seen, with hindsight, that the introduction of the computer into the humanities disciplines has made no more substantial change in the

eventual output of research than was made by the introduction of the electric typewriter? Or will there be a flowering of research papers, neither pedestrian nor spectacular, containing solid and original results obtained by techniques for which the computer is indispensable? It is hard to make a confident prediction. But the testing time has now arrived; because for the first time posts of leadership in humanities departments are being taken up by a generation of scholars who have been familiar with the computer . . .

Although the use of computers in humanities dates back to the 1960s, this extract from a published lecture shows the state of development in the early 1990s, 30 years later. Through analysis, the BNC can provide an understanding of the meanings, collocations and grammatical patterns of *humanity* (and *humanities*) in a large general English corpus from the last century. These two analyses begin to illuminate the term, and much more could be done with further exploration of a wider range of uses and more sophisticated techniques.

BAWE

The British Academic Written English (BAWE) corpus allows us to search for *humanity* by genre family, level of study and discipline to find out how it is used in student writing in the early twenty-first century. If we filter to discipline and search for collocates, we see that it collocates differently in law, philosophy/ethics, English literature and classics, as in these examples.

'crimes against humanity' in law

individuals at the highest levels of government and military infrastructure could be prosecuted and punished under international law, recognising accountability for the commission of war crimes and introduced new offences of crimes against peace and crimes against humanity.

'common humanity' and 'humanity can survive' in ethics/philosophy

Indeed it is the similarities between so-called cultures that allow us to appreciate that there is a common humanity. If there was not then we would be unable to comment on the cultures of the past or of different location because there would be nothing we could relate to. This is shown in . . .

'his/her humanity' in English literature

a microcosmic happy ending in *Emma* prefigures the macrocosmic happy ending of the whole book, for Mr Knightley, being true to his name, moderates Harriet's embarrassment by dancing with her himself, thus proving his humanity and that he is the ideal husband for Emma.

Nineteen of the 31 instances of 'humanity' in the law texts relate to crime(s) against humanity, a phrase which is also found in politics and history, though less frequently. In contrast, philosophy is concerned with humanity in general, with our 'common humanity' and how 'humanity can survive' together, accounting for 12 of the 28 instances of 'humanity' in the philosophy texts where it has a wider range of uses, including concern for the survival/the fall/the future of humanity. In English it is used in relation to specific individuals and their

Table 3.2 Concordance lines of ‘humanity’ in the discipline of classics in the BAWE corpus

Classics	As the play goes on we watch as Medea’s	humanity	falls away.
Classics	effectively shutting off the last of her	humanity	as, when
Classics	argue against the games, but not for the case of	humanity	. Instead, these members
Classics	really a lady who does not possess any	humanity	Representing
Classics	in the conventional sense, involves loss of	humanity	’. The final sacrifice
Classics	good of the future Rome, and so can submit to his	humanity	. This is not to say however,
Classics	of Caesar, although he tries to diminish his	humanity	Caesar had

relation to acts of human kindness, as in the example of Mr Knightley, while in classics it tends to have more negative associations around the loss of humanity, as shown in Table 3.2.

Although the numbers of instances are small, such disciplinary variations in form, meaning and use are characteristic of academic language. In contrast with the BNC, expressions such as ‘dregs of humanity’ are not found.

COCA

A search in the BYU Corpus of Contemporary American English (COCA) includes five ‘genres’: spoken, fiction, popular magazine, newspaper and academic, evenly balanced for size over five-year spans from 1990 to 2015. Here ‘humanity’ is used least (10.13 pm) in spoken texts and most (34.85 pm) in academic journals.

Among the academic disciplines, ‘humanity’ is used extensively in philosophy and religious studies (179.37 pm), an average amount in humanities (35.0 5pm), and perhaps surprisingly it is seldom used in education (8.05 pm) and medicine (1.79 pm). Further investigation might explore whether these are consistent characteristics of the disciplines or accidental findings reflecting topics in the articles selected for the corpus.

The top collocates include *against*, *crime(s)*, *common* and *mass*. Interestingly, the second most frequent collocate is *habitat*, which introduces ‘Habitat for Humanity’, which occurs in newspapers rather than in academic texts.

Give just \$35 to **Habitat** for **Humanity** at habitat.org/haiti to help repair one damaged home in Port-au-Prince

here’s a great idea: Volunteer with **Habitat** for **Humanity**.

A good way of recycling cabinets and appliances is to donate them to **Habitat** for **Humanity** or repurpose them . . .

This phrase is also found in one news item in the BNC, but, unsurprisingly, not in the BAWE.

‘Humanity’ occurs throughout the decades in COCA, and most frequently between 2005 and 2009, where it is associated with *crimes against humanity*, such as genocide or violence against innocent civilians. The examples point to this being a period of news reporting and academic discussion of crimes against humanity in countries such as Sudan, Zimbabwe, the former Yugoslavia, Angola and Iraq.

new

Corpus of Contemporary American English





SEARCH

FREQUENCY

CONTEXT

INTERFACE

SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...]

COMPARE

	☐	CONTEXT	FREQ		ALL	%	MI	
1	☐	AGAINST	798		263709	0.30	4.21	
2	☐	CRIMES	669		15705	4.26	8.03	
3	☒	HABITAT	383		9888	3.87	7.89	
4	☐	GOD	271		121865	0.22	3.77	
5	☐	COMMON	238		80554	0.30	4.18	
6	☐	CRIME	137		45581	0.30	4.20	
7	☐	NATURE	127		73813	0.17	3.40	
8	☐	CHRIST	90		21293	0.42	4.70	
9	☐	JESUS	89		29957	0.30	4.19	
10	☐	GENOCIDE	68		3631	1.87	6.84	

Figure 3.1 Frequency and MI of top ten collocates of ‘humanity’ in COCA

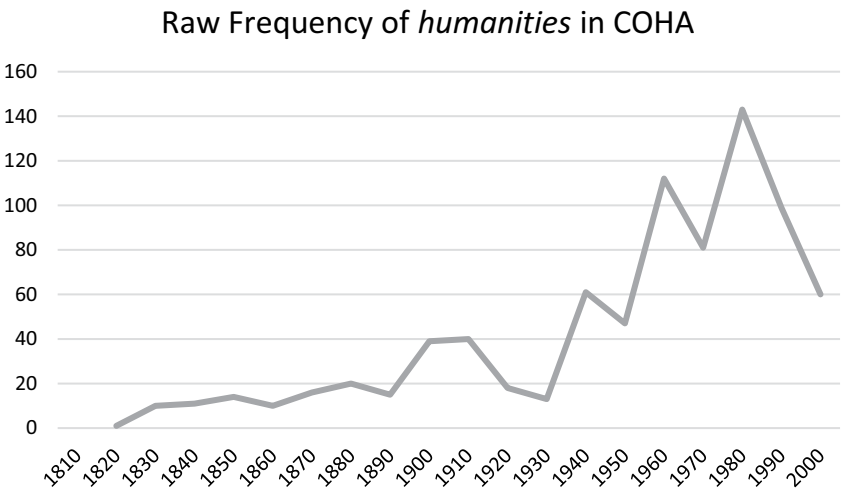


Figure 3.2 Raw frequency of ‘humanities’ in COHA

Further searches show that *humanity* in COCA has four main uses: crimes against humanity (see *crimes*, *crime*, *against* and *genocide* in Figure 3.1), Habitat for Humanity (3.87 percent of the instances of *habitat* occur with *humanity* with 7.89 mutual information (MI), which indicates the strength of the collocation), *humanity* in contrast with God (*God* accounts for 0.22 percent with 3.77 MI), and *humanity* meaning compassion, while *humanities* in COCA is associated with features of American university culture, such as *college*, *endowment* and *majors*.

COHA

A search in The Corpus of Historical American English (COHA) suggests that *humanities* is most frequent in the latter part of the twentieth century, occurring (Figure 3.2) more than 80 times in the 1960s–1990s, compared to fewer than 20 times in the 1820s–1890s.

Earlier examples, as this one from the 1840s, are invoked in relation to education:

this process might be called humanization, i. e., the complete drawing out and unfolding of his proper nature making him perfectly a man-realizing his ideal character; (and hence, with singularly beautiful appropriateness, the proper studies of a liberal education used to be called, not only the Arts, but the Humanities).

The change in meaning referred to is also reflected in the relatively strong and enduring collocate ‘so-called’ which suggests *humanities* is not an established term (Table 3.3).

The strongest collocates at the end of the twentieth century (*science, arts, social, national, endowment*) are similar in many respects to those in Figure 3.3, reflecting a fully institutionalised meaning of humanities alongside arts and social sciences with departments, professors, and endowments. No collocates are found, however, of *digital* or *research*, suggesting these came later.

Hansard

Hansard is a transcribed record of speeches and debate in the UK House of Commons and House of Lords (<https://hansard.parliament.uk/about>, accessed 17/04/2018). A search for the term ‘humanities’ in the Hansard corpus points to the debate over a Humanities Research Council in the 1990s where we see the humanities decried as a ‘waste of time’ and where ‘so-called humanities’ appears pejoratively.

GloWbE

The various digital humanities movements perhaps aim to establish a clear, contemporary identity for humanities, and a search in the 1.9-billion-word GloWbE corpus of web material from 20 English-speaking countries in 2012–2013 reveals that the term *digital humanities*

Table 3.3 ‘So-called humanities’ in COCA

1887	MAG	. . . courses of study, in one of which letters and the so-called humanities . . .
1927	NF	. . . the so-called Humanities, or classics of Ancient Greece and Rome . . .
1975	FIC	. . . - decadence of art, scientific endeavour, the so-called humanities -
2004	NF	. . . is especially though not exclusively the case in the so-called humanities . . .

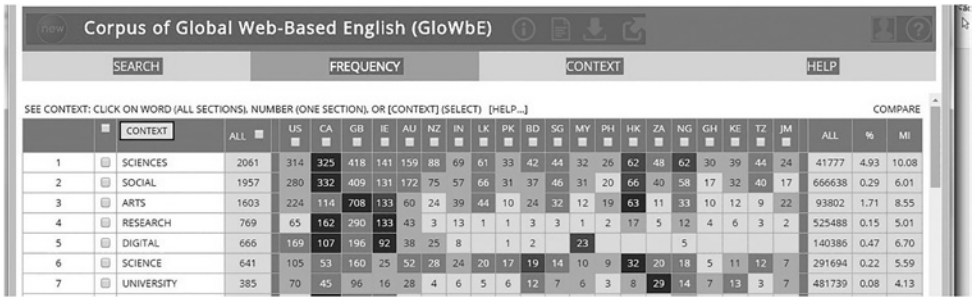


Figure 3.3 Collocates of ‘humanities’ in GloWbE

appears on the Web extensively in the UK, Canada, Australia and New Zealand, but to a lesser extent in India, Malaysia and Pakistan, and not at all in Hong Kong, Jamaica or African countries such as Ghana, Kenya, Tanzania and South Africa (Figure 3.3).

The one GloWbE instance of ‘so-called humanities’ is from an American site, but refers to Europe:

The European educational system’s current process of so-called reform is marked by de-financing, cuts, job losses, and also by a downsizing of non-rentable disciplinary fields – the **so-called humanities** – accompanied by the increased support of capital-intensive fields of research. The leading principle of this reform is the assertion of the epistemological primacy of the economic sphere.

These searches for *humanities* show that while corpus investigations can produce statistical information about frequencies and collocations, it is important that users pay heed to the nature of the data in the corpus to inform interpretations. An understanding of the nature of the texts in the corpus, the contexts of their production and how they were selected all play a role in interpreting the significance of findings. We have glimpsed how pre-loaded corpora can be used to demonstrate the range of grammatical patterns in large general corpora associated with search items (through WordSketches), to suggest lines of enquiry (e.g. to what extent is *humanities* a contested term?) and to compare data from comparable sources (e.g. from different national web domains or from student writing in law vs. literature). To answer specific research questions using written corpora, however, requires careful attention to corpus selection, preparation and interrogation.

Small correspondence corpora

This study uses three small correspondence corpora to explore some of the linguistic features that are typically found in letters. The *Migrant Correspondence Corpus* (MCC) contains 188 private letters (176,501 tokens) between Irish migrants and their families, dating from 1819 to 1953; roughly 75 percent of the letters are by male authors (141 letters) and 25 percent (47 letters) by female authors. The *British Telecom Correspondence Corpus* (BTCC) contains 612 business letters (130,000 tokens) dating from 1853 to 1982. The letters – sourced from the public archives of BT – are predominately written by men, with only 13 authors identifying themselves as female. Finally, the *Letters of Artisans and the Labouring Poor* (LALP) corpus contains 272 letters (51,408 tokens) dating from 1720 to 1837 written by individuals appealing for poor relief from their parish or poor law authority; 118 male and 66 female authors are represented in the corpus. For this study letters from Record Offices in Warwickshire, Lancashire and Essex were used. The corpora represent different types of correspondence (personal and professional letters, as well as letters of solicitation), allowing users to explore the discourse of correspondence from a diachronic perspective. This analysis will investigate the register of correspondence, examining the extent to which the type of letter (personal or professional, for instance) predicts variation in the language. It will also explore how the MCC, BTCC and LALP data compare to modern English language registers of other kinds.⁴

The MCC, BTCC and LALP letters are all saved in XML – a format that is compatible with most corpus software. Each collection has an accompanying spreadsheet containing metadata relating to the individual XML letter files. Metadata provides additional information or knowledge about the text. It can be situated within the body of the text, providing

information or knowledge relating to the structure, layout or content (line breaks, paragraphs or pragmatic features, for instance); and it can be situated outside the body of the text – in the header – providing information or knowledge about where the text comes from (bibliographic information) and what it is (a letter, an academic essay or a poem, for instance). Most corpora will contain some basic header information; however, the type of information captured will vary from project to project depending on the research aims. While the MCC header information includes extensive sociobiographic details about the various participants (their sex, occupation, educational background and family history), BTCC header information focuses more on the professional status of the various authors and recipients and the communicative function of the letter itself. The LALP corpus, in contrast, captures header information about the materiality of the letter – whether it was written in pen or pencil and the quality of the paper, as well as details about the authenticity and authorship of the correspondence (whether or not the letter was dictated, for example).

Spreadsheets and databases are ideal for capturing well-structured metadata such as the header information described earlier. However, to use this metadata with corpus tools, it must be encoded in a way that the software will recognise. While there are many ways of encoding metadata, the *Text Encoding Initiative* (TEI) is the de facto standard for encoding digitised texts in the humanities. Although many of the information categories used in the MCC, BTCC and LALP corpus are varied, they do share some common features: a sender, a recipient, an origin, a destination and a date. The TEI has recently introduced a module for capturing this information in the <profileDesc> section of the TEI header (see Figure 3.4), allowing different letter collections to interconnect based on these shared attributes (see Stadler et al. 2016). Provided the information categories are TEI compatible, it is relatively straightforward from a programming perspective for metadata to be extracted from the spreadsheet and rendered as TEI headers within the corresponding XML files. These XML files can then be uploaded into corpus software, allowing the user to refine their search queries based on the header information (searching for all letters from a particular period or location, or by a particular author, for instance).

```
<profileDesc>
  <correspDesc xmlns="http://wiki.tei-c.org/index.php/SIG:Correspondence/
task-force-correspDesc">
    <correspAction type="sending">
      <persName>Elizabeth Lough</persName>
      <settlement>Winsted, Connecticut</settlement>
      <date when="1876-03-07"/>
    </correspAction>
    <correspAction type="receiving">
      <persName>Elizabeth McDonald Lough</persName>
      <settlement>Meelick, Queen's County</settlement>
    </correspAction>
  </correspDesc>
</profileDesc>
```

Figure 3.4 TEI encoding for capturing information about the sender, recipient, origin, destination and date, using metadata from the MCC as an example

To get a sense of how the letters in the MCC, BTCC and LALP corpus compare with other English language registers, the data were uploaded into Nini's (2015) *Multidimensional Analysis Tagger* (MAT). MAT draws on Biber (1988) by grammatically annotating the corpus and plotting the distribution of corpus holdings based on Biber's (1988/1995) dimensions. Dimension 1, 'Informational versus Involved Production', represents 'a fundamental parameter of variation among texts in English' (Biber 1988: 115). While one end of the pole represents 'discourse with interaction, affective, involved purposes' (telephone or face-to-face conversations, for example), the other end represents 'discourse with highly informational purposes' (official documents, or academic prose etc.) (ibid.). In terms of correspondence, Biber's study showed that personal letters had a mean score of 19.5 for dimension 1 (placing them alongside spontaneous speeches and interviews), and professional letters had a mean score of -3.9 (placing them alongside general fiction and broadcasts). In other words, although both types of correspondence are interactional (as evidenced by the high frequency of first- and second-person pronouns), they differ in terms of their functions: while personal letters are more interpersonal, sharing many of the same characteristics as conversations, professional letters are more informational, sharing characteristics typically found in written discourse. Table 3.4 compares the MCC, BTCC and LALP corpus with Biber's mean scores for personal letters and professional letters.

Similar to Biber's findings for professional letters, Table 3.4 shows that the BTCC letters have a mean score of -7.1 for dimension 1, placing them at the 'informational' end of the 'Informational versus Involved Production' spectrum. However, unlike personal letters with a score of 19.5, the MCC letters have a mean score of 0.36, placing them alongside general fiction. In other words, the migrant letters are much less 'affective' and 'involved' than the personal letters used for Biber's study. The LALP letters are quite interesting; with a mean score of 2.17, they are more conversation-like in nature than the MCC letters and correspond most closely with Biber's prepared speeches.

Having looked at the dimension scores for the three corpora, the next stage involved examining some of the linguistic features that are common across all three datasets: namely, first- and second-person pronouns and what Biber describes as private verbs – that is, verbs which 'express intellectual states or nonobservable intellectual acts' (ibid.) – typically verbs of cognition such as 'I *think*' or 'she *hopes*'. In systemic-functional grammar (e.g. Halliday and Matthiessen 2004) these clauses can function as a type of projection. Projection structures consist of two main components: the projecting clause (*I hope*) and the projected clause (*you will write*). In these structures the primary (projecting) clause sets up the

Table 3.4 Mean scores for dimensions across five datasets

	<i>Personal Letters</i>	<i>Professional Letters</i>	<i>MCC</i>	<i>BTCC</i>	<i>LALP</i>
Dimension 1	19.5	-3.9	0.36	-7.1	2.17
Dimension 2	0.3	-2.2	-0.46	-1.49	-1.21
Dimension 3	-3.6	6.5	1.46	5.15	2.02
Dimension 4	1.5	3.5	1.75	6.69	6.88
Dimension 5	-2.8	0.4	0.34	2.59	0.09
Dimension 6	-1.4	1.5	-0.22	1.64	-0.5

secondary (projected) clause as the representation of the content of either what is thought or what is said (Halliday and Matthiessen 2004: 377). Halliday and Matthiessen make a distinction between the projection of propositions and the projection of proposals as follows:

Propositions, which are exchanges of information [typically statements or questions], are projected mentally by processes of cognition – thinking, knowing, understanding, wondering, etc. . . . Proposals, which are exchanges of goods-&-services [typically offers or commands], are projected mentally by processes of desire.

(2004: 461)

Both propositions and proposals have different response-expecting speech functions, and an analysis of these structures will reveal something about how the author interacts with their intended recipient and the type of response they expect – whether that is a verbal response (to provide information) or a non-verbal response (to carry out an action).

Sketch Engine (Kilgariff and Kosem 2012) was used to automatically tag the corpus data for part of speech (POS), using the Penn Treebank tagset, making it possible to use Corpus Query Language (CQL) to identify lexico-grammatical patterns within each dataset. The following patterns were extracted (variants, e.g. with other pronouns or intervening adverbs, are not considered here.):

- [tag=“PP”] to search for all personal pronouns
- [word=“I”][tag=“V.”] to search for all personal pronoun + verb combinations
- [word=“I”][tag=“V.”][word=“you”] to search for all personal pronoun + verb + personal pronoun combinations

Looking at the normalised frequencies (averaged per million), Table 3.5 shows similar frequencies for personal pronouns in the MCC and LALP (87,794 and 84,033, respectively); however, there are significantly fewer in the BTCC (51,113). It is a similar trend for the patterns *I + Verb* and *I + Verb + You*.

Having looked at the frequency of the pattern *I + V + you*, the next stage was to see which verbs typically appear in this structure (see Table 3.6). Not all of the examples in Table 3.6 are types of projection. *I thank you*, for instance, is a straightforward subject/verb/object construction. The most common verbs of desire (which typically realise the projection of proposals) are *hope* and *wish*, occurring across all three corpora. The verb *want* also occurs in the MCC and BTCC. The most common verbs of cognition (which typically realise the projection of propositions) are *think* (across all three corpora), *suppose* (in the MCC and BTCC) and *know* (which has a high frequency in the MCC, but only occurs once in the BTCC and not at all in the LALP corpus). Interestingly, the LALP corpus contains

Table 3.5 Raw and normalised frequencies (per million) for Pronouns, I + Verb, and I + Verb + You across the MCC, BTCC and LALP corpus

	MCC	BTCC	LALP
Pr	16,814 (87,794 p/m)	7,511 (51,113 p/m)	4,332 (84,033 p/m)
I + V	3,951 (20,630 p/m)	1,821 (12,326 p/m)	912 (17,728 p/m)
I + V + you	212 (1,106 p/m)	73 (494 p/m)	48 (933.05 p/m)

a high frequency of verbs of desire (*hope, wish, desire*), but only one verb of cognition (*think*); additionally, there are several instances of what Biber (following Quirk et al. 1985: 1182–1183) describes as suasive verbs (verbs that ‘imply intentions to bring about some change in the future’ (1988: 242) – *beg, solicit* and *ask*).

I hope you is the most frequent projection structure across all three corpora; however, a closer examination of the language in context revealed some differences. In the MCC, *I hope you* is typically used in formulaic greetings (21 out of 37 occurrences – see concordance lines 1 and 2 in Figure 3.5, for instance). In the remaining 16 occurrences, the recipient of the letter (*you*) is typically required to *think, pray*, or *remember/not forget* (see concordance lines 3 to 6). In the BTCC only 3 out of 13 instances of *I hope you* are part of a formulaic greeting. In over half of the occurrences, what follows are the modals *may, will* or *would* + *agree* (see concordance lines 7 to 9). Finally, in the LALP corpus, 32 of the 39 occurrences of *I hope you* are followed by *will*. Here the recipient is typically required to *send* money (concordance line 10), *speak* or *write* to somebody on the author’s behalf (concordance line 11) or *consider* the author’s request (concordance line 12).

What is significant about projection structures is their ability to directly address and involve the recipient of the letter, assigning to them a role to play in the communicative event that is taking place. In the MCC, the subject of the projected clause – *you* (typically a close family member) – is required to respond cognitively in some way, by remembering a relative or by not feeling abandoned or forgotten. Similarly, in the BTCC data, for the most part, the subject of the projected clause (a business associate) is also required to respond cognitively – this time, by agreeing to something. In contrast, the subject of the projected clause in the LALP data (the poor law authority) is almost always required to carry out a

Table 3.6 Instances of I + Verb + You across the three corpora, organised by frequency

	MCC I + V + You	BTCC I + V + You	LALP I + V + You
1	I hope you (61)	I thank you (15)	I hope you (25)
2	I tell you (16)	I hope you (13)	I trust you (3)
3	I wish you (13)	I told you (6)	I wish you (3)
4	I suppose you (12)	I think you (5)	I thank you (2)
5	I know you (12)	I sent you (5)	I beg you (2)
6	I want you (11)	I wish you (3)	I desire you (2)
7	I told you (8)	I suppose you (3)	I think you (1)
8	I think you (8)	I send you (2)	I solicit you (1)
9	I send you (6)	I see you (2)	I sent you (1)
10	I left you (6)	I wrote you (1)	I return you (1)
11	I sent you (4)	I wanted you (1)	I refer you (1)
12	I wrote you (3)	I trust you (1)	I ask you (1)
13	I expect you (3)	I trouble you (1)	
14	I assure you (3)	I thought you (1)	
15	I write you (2)	I telegraphed you (1)	
16	I thought you (2)	I reminded you (1)	
17	I see you (2)	I promise you (1)	
18	I saw you (2)	I offer you (1)	
19	I remain you (2)	I observe you (1)	
20	I refer you (2)	I know you (1)	

1) I was not one day sick since I left you ____ I hope you	and the Children is well and that the Cap was in good health
2) I am now in good health thank God ____ I hope you	are well and all friends in Ireland Remember to all friends
3) all are and how you are getting along ____ I hope you	dont think I have lost all interest in you or in what concerns
4) June 25th 1855 My Dearest Mother ____ I hope you	do not think that I have forgotten you by My not answering
5) & morning in my Poor Prayers and ____ I hope you	wonte forget your Old Uncle well I will Come to a Close by
6) s Suport the poor – My Dear terese ____ I hope you	will pray often for thy Grandfather and have his name inserte
7) would otherwise have been assigned. I hope you	will agree, and that we may rely on your co-operation in appl
8) ce a letter of which I enclose a copy. I hope you	will agree generally with this assessment. Please let us know
9) matters would not come to that point. I hope you	would agree that a situation in which BT was force to permit
10) this morning only it Was so wet so ____ I hope you	will send me something by return of post Benjamin Hewitt
11) do not a lowe mesomething more and I hope you	will be so good as to speeke to thee gentleman a bout me
12) Myself have Sufficient Nesseseries ____ I hope you	will Consider of my Circumstance for I have done as far as

Figure 3.5 Sample concordance lines for ‘I hope you’

physical action – to write or send money, or to *consider* carrying out these actions. It would seem that certain projection structures (for example, *I hope you*) may be indicative of the letter writing genre; however, their use varies depending on the type of letter it is (personal, professional, solicitation). Additionally, the author–recipient relationship would certainly influence language choice. Moreton (2015), for instance, examining a collection of migrant correspondence, found that letters to siblings, nieces or nephews (i.e. an ‘inferior’ within the notional familial hierarchy) contained more projections of proposals (often realising indirect commands), while letters to parents (i.e. a generational ‘superior’) contained more projections of propositions (typically statements/exchanges of information). Similarly, Sairio’s study of letters by Samuel Johnson to two of his correspondents (Mrs Thrale, a close friend, and Lucy Porter, Johnson’s step-daughter) found that the use of first- and second-person pronouns as well as evidential verbs (such as *know*, *think* and *believe*) are ‘a relevant indicator of the closeness of the relationship’ (2005: 33): the closer the relationship, the more likely it is that these linguistic devices will be used.

Future directions

Despite the exciting opportunities offered by the digital humanities with regard to digitising written resources, several constraints may hinder future research. Differing laws and attitudes relating to copyright and intellectual property across disciplines, cultures and countries affect accessibility of resources and are among the barriers to creating and interconnecting corpora. Certainly, when working with historical manuscripts, the only way around the problem, as argued by Honkapohja et al., is to ‘work with repositories and persuade them to either digitise the manuscript material and to publish them under an open-access licence, or to allow scholars to photograph manuscript material themselves’ (2009: 10). Sustainability of digital resources is also an issue, which Millett argues is exacerbated, in part, by ‘the tendency of large electronic projects . . . to use complex custom-built software, which makes them particularly difficult to update’ (2013: 46). And, finally, undocumented and/or differing transcription and encoding practices can lead to the duplication of work and make data interchange more challenging. However, alliances between disciplines are starting to form, largely driven by the need for reliable digital transcriptions of manuscripts that can be used across a range of disciplines. The *Digital Editions for Corpus Linguistics* (DECL) project, for example, ‘aims to create a framework for producing online editions of historical manuscripts suitable for both corpus linguistic

and historical research' (Honkapohja et al. 2009: 451). Ultimately, what is needed is a cross-disciplinary, cross-cultural and cross-institutional approach to digitisation and annotation of corpus data.

Notes

- 1 University of Oxford (2009) *Cultures of knowledge*. Available at: www.culturesofknowledge.org.
Stanford University (2013) *Mapping the republic of letters*. Available at: <http://republicofletters.stanford.edu/index.html>.
University of Oxford (2008) *Electronic Enlightenment*. Available at: www.e-enlightenment.com/info/about/.
Cambridge University (2015) *The darwin project*. Available at: www.darwinproject.ac.uk/darwins-letters.
University of South Carolina (2011) *Victorian lives and letters consortium*. Available at: <http://tundra.csd.sc.edu/vllc/>.
- 2 Baylor University. *The browning letters*. Available at: <http://digitalcollections.baylor.edu/cdm/landingpage/collection/ab-letters>.
University of Glasgow. *Bess of Hardwick's letters: The complete correspondence c.1550–1608*. Available at: www.bessofhardwick.org.
Van Gogh Museum. Vincent van Gogh: The Letters. Available at: <http://vangoghletters.org/vg/>.
- 3 A term coined by the Spanish writer Miguel de Unamuno in 1985. 'It refers to the value of the humble and anonymous lives experienced by ordinary men and women in everyday contexts which form the essence of normal social interactions, as opposed to the lives of leaders and famous people that are generally accounted for in canonical histories' (Amador-Moreno et al. 2016).
- 4 The letters used to create the MCC are from Professor Kerby Miller's archive of Irish migrant correspondence, housed at the University of Missouri. Miller has published extensively on the topic of Irish migration, see, for e.g. Miller 1985; Miller et al. 2003. The authors would like to thank Professor Miller for providing access to the letters. The BTCC data can be accessed from: <https://btletters.wordpress.com>. For more information about the LALP project see <https://lalpcorpus.wordpress.com/>.

Further reading

- 1 *Creating and digitizing language corpora* series (Volumes 1–3), published by Palgrave Macmillan. This is a useful source of information about corpus design and creation.
- 2 Nesi, H. (2016). Corpus studies in EAP. In K. Hyland and P. Shaw (eds.), *Routledge handbook of English for academic purposes*. Abingdon: Routledge, pp. 206–217.

This provides an overview of corpora, both written and spoken, in English for academic purposes.

More information about written corpora and related publications can be found at:

- The Oxford Text Archive (OTA) <http://ota.ox.ac.uk>
- The Learner Corpora lists www.uclouvain.be/en-cecl-lcworld.html
- The Corpus Resource Database (CoRD) www.helsinki.fi/varieng/CoRD/
- The Sketch Engine site, www.sketchengine.co.uk
- The Brigham Young University site, <http://corpus.byu.edu>

The OTA houses dozens of corpora that can be requested by researchers in a range of formats. In contrast, the Sketch Engine and BYU sites can be used for researcher-compiled corpora, and both include dozens of pre-loaded corpora that can be easily analysed by novices.

The Sketch Engine site developed initially by the late Adam Kilgarriff includes the British National Corpus and the American National Corpus; the Brown Corpus, the Literary corpora of Early English Books Online and Project Gutenberg; web corpora such as Web 2008, 2012, 2013, uWAC (British web corpus) and Wikipedia; as well as more specialist corpora that will be of interest to researchers in the humanities, such as web corpora of African and Asian English (Araneum

Anglicum), the CHILDS corpus, the SiBOL/port newspaper corpus, a corpus of law reports, of international art (e-flux) and academic English (BAWE).

The BYU site developed by Mark Davies includes GloWbE, the NOW corpus and a Corpus of Online Registers of English (CORE). In addition to two corpora of British English (the BNC and Hansard) and one corpus of Canadian English, there are four corpora of American English, including COCA, COHA, Time Magazine and contemporary soap operas.

References

- Amador-Moreno, C., Corrigan, K.P., McCafferty, K., Moreton, E. and Waters, C. (2016). Irish migration databases as impact tools in the education and heritage sectors. In K. Corrigan and A. Mearnes (eds.), *Creating and digitizing language corpora – volume 3: Databases for public engagement*. London: Palgrave Macmillan.
- Auer, A. and Fairman, T. (2012). Letters of artisans and the labouring poor (England, c. 1750–1835). In P. Bennett, M. Durrell, S. Scheible and R.J. Whitt (eds.), *New methods in historical corpus linguistics*. Narr: Tübingen, pp. 77–91.
- Baron, A., Tagg, C., Rayson, P., Greenwood, P., Walkerdine, J. and Rashid, A. (2011). Using verifiable author data: Gender and spelling differences in Twitter and SMS. Presented at *ICAME 32*, Oslo, Norway, 1–5 June.
- Biber, D. (1988/1995). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D., Conrad, S., Reppen, R., Byrd, P. and Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly* 36(1): 9–48.
- Biber, D., Finegan, E. and Atkinson, D. (1994). ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers. In U. Fries, P. Schneider and G. Tottie (eds.), *Creating and using English language corpora: Papers from the 14th international conference on English language research on computerized corpora, Zurich 1993*. Amsterdam: Rodopi.
- Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Pearson Education.
- Caldas-Coulthard, C.R. and Moon, R. (2010). Curvy, hunky, kinky’: Using corpora as tools for critical analysis. *Discourse & Society* 21(2): 99–133.
- Durrant, P. (2017). Lexical bundles and disciplinary variation in university students’ writing: Mapping the territories. *Applied Linguistics* 38(2): 165–193.
- Fairman, T. (2012). Letters in mechanically-schooled language. In M. Dossena and G. Del Lungo Camiciotti (eds.), *Letter writing in late modern Europe*. Amsterdam: John Benjamins Publishing Company, pp. 205–227.
- Fairman, T. (2015). Language in print and handwriting. In A. Auer, D. Schreier and R.J. Watts (eds.), *Letter writing and language change*. Cambridge: Cambridge University Press, pp. 53–71.
- Fludernik, M. (1991). Shifters and deixis: Some reflections on Jakobson, Jespersen, and reference. *Semiotica* 86: 193–230. <https://doi.org/10.1515/semi.1991.86.3-4.193>.
- Francis, G., Hunston, S. and Manning, E. (1996). *Grammar patterns 1: Verbs*. London: HarperCollins.
- Frankenberg-Garcia, A., Flowerdew, L. and Aston, G. (2011). *New trends in corpora and language learning*. London: Continuum.
- Fries, P.H. (2010). Charles C. Fries, linguistics and corpus linguistics. *ICAME Journal: Computers in English Linguistics* 34: 89–121.
- Gardner, S. and Holmes, J. (2009). *Can I use headings in my essay?* Section headings, macrostructures and genre families in the BAWE corpus of student writing. In M. Charles, S. Hunston and D. Pecorari (eds.), *Academic writing: At the interface of corpus and discourse*. London: Continuum, pp. 251–271.
- Gries, S.T. (2013). *Statistics for linguistics with R: A practical introduction* (2nd ed.). Berlin: Walter de Gruyter.
- Halliday, M.A.K. and Matthiessen, C.M.I.M. (2004). *An introduction to functional grammar*. London: Arnold.

- Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Abingdon and New York: Routledge.
- Holmes, J. and Nesi, H. (2009). Verbal and mental processes in academic disciplines. In M. Charles, S. Hunston and D. Pecorari (eds.), *Academic writing: At the interface of corpus and discourse*. London: Continuum, pp. 58–72.
- Honkapohja, A., Kaislaniemi, S. and Marttila, V. (2009). Digital editions for corpus linguistics: Representing manuscript reality in electronic corpora. In A.H. Jucker, D. Schreier and M. Hundt (eds.), *Corpora: Pragmatics and discourse: Papers from the 29th international conference on English language research on computerized corpora (ICAME 29)*. Amsterdam and New York: Rodopi.
- Hornby, A.S. (1974). *Oxford advanced learner's dictionary* (3rd ed.). Oxford: Oxford University Press.
- Huber, M. and Mukherjee, J. (2013). 'Introduction' to corpus linguistics and variation in English: Focus on non-native Englishes. *Varieng: Studies in Variation, Contacts and Change in English*: 13.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London: Continuum.
- Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27: 4–21.
- Kilgariff, A. and Kosem, I. (2012). Corpus tools for lexicographers. In S. Granger and M. Paquot (eds.), *Electronic lexicography*. New York: Oxford University Press, pp. 31–56.
- Lea, D. (2014). *Oxford learner's dictionary of academic English*. Oxford: Oxford University Press.
- Leedham, M. (2015). *Chinese students' writing in English: Implications from a corpus-driven study*. Abingdon: Routledge.
- Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes* 31(1): 25–35.
- Miller, K.A. (1985). *Emigrants and exiles: Ireland and the Irish Exodus to North America*. New York: Oxford University Press.
- Miller, K.A., Schrier, A., Boling, B.D. and Doyle, D.N. (2003). *Irish immigrants in the land of Canaan: Letters and memoirs from colonial and revolutionary America, 1675–1815*. New York: Oxford University Press.
- Millett, B. (2013). Whatever happened to electronic editing? In V. Gillespie and A. Hudson (eds.), *Probable truth: Editing medieval texts from Britain in the twenty-first century*. Turnhout: Brepols, pp. 39–54.
- Moon, R. (2007). Sinclair, lexicography, and the Cobuild project: The application of theory. *International Journal of Corpus Linguistics* 12(2): 159–181.
- Moreton, E. (2015). I hope you will write: The function of projection structures in a corpus of nineteenth century Irish emigrant correspondence. *Journal of Historical Pragmatics* 16(1): 277–303. doi:10.1075/jhp.16.2.06mor.
- Nesi, H. (2016). Corpus studies in EAP. In K. Hyland and P. Shaw (eds.), *The Routledge handbook of English for academic purposes*. Abingdon: Routledge, pp. 206–217.
- Nesi, H. and Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. Applied Linguistics Series. Cambridge: Cambridge University Press.
- Nesi, H. and Moreton, E. (2012). EFL/ESL writers and the use of shell nouns. In R. Tang (ed.), *Academic writing in a second or foreign language: Issues and challenges facing ESL/EFL academic writers in higher education contexts* (pp. 126–145). London: Continuum.
- Nini, A. (2015). *Multidimensional analysis tagger* (Version 1.3). Available at: <http://sites.google.com/site/multidimensionaltagger>.
- Preston, M. (n.d.). *A brief history of computer concordances*. Available at: www.colorado.edu/Arts_Sciences/CCRH/history.html (Accessed 15 December 2016).
- Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.
- Raumolin-Brunberg, H. and Nevalainen, T. (2007). Historical sociolinguistics: The corpus of early English correspondence. In J. Beal, K. Corrigan and H. Moisl (eds.), *Creating and digitizing language corpora* (Vol. 2). London: Palgrave Macmillan.

- Rayson, P. and Baron, A. (2011). Automatic error tagging of spelling mistakes in learner corpora. In F. Meunier, S. De Cock, G. Gilquin and M. Paquot (eds.), *A taste for corpora. In honour of Sylviane granger*. Studies in Corpus Linguistics, 45. Amsterdam: John Benjamins.
- Sairio, A. (2005). ‘Sam of Streatham Park’: A linguistic study of Dr. Johnson’s membership in the Thrale family. *European Journal of English Studies* 9(1): 21–35.
- Sairio, A. (2009). *Language and letters of the bluestocking network. Sociolinguistic issues in eighteenth-century epistolary English* (Monograph). (Mémoires de la Société Néophilologique de Helsinki 75). Helsinki: Société Néophilologique.
- Sifton, P.G. (1977). The provenance of the Thomas Jefferson papers. *American Archivist* 40(1): 17–30.
- Stadler, P., Illetschko, M. and Seifert, S. (2016). Towards a model for encoding correspondence in the TEI: Developing and implementing <correspDesc>. *Journal of the Text Encoding Initiative* (9). Available at: <http://jtei.revues.org/1433>.
- Staples, S., Egbert, J., Biber, D. and Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication* 33(2): 149–183.
- Tagg, C., Baron, A. and Rayson, P. (2012). “i didn’t spel that wrong did i. Oops”: Analysis and normalisation of SMS spelling variation. In L.A. Coughon and C. Fairon (eds.), *SMS communication: A linguistic approach*. Special issue of *Linguisticae Investigationes* 35(2): pp. 367–388.
- Thompson, G. and Hunston, S. (2006). System and corpus: Two traditions with a common ground. In G. Thompson and S. Hunston (eds.), *System and corpus: Exploring connections*. London: Equinox, pp. 1–14.
- Tieken-Boon van Ostade, I.M. (2010). *The Bishop’s grammar: Robert Lowth and the rise of prescriptivism*. Oxford: Oxford University Press.
- Willis, D. (1994). A lexical approach. In M. Bygate, A. Tonkyn and E. Williams (eds.), *Grammar and the language teacher*. Hemel Hempstead: Prentice Hall International.
- Wisby, R. (1962). Concordance making by electronic computer: Some experiences with the wiener genesis. *The Modern Language Review* 57(2): 161–172. doi:10.2307/3720960.

Digital interaction

Jai Mackenzie

Introduction

The question of how digital technologies can both shed light on and affect the human experience (and indeed, how the human experience affects digital technologies) is an abiding concern for research in the digital humanities (Burdick et al. 2012; Evans and Rees 2012). Research at the intersection of English language and the digital humanities is particularly well situated to answer this question and has revealed much about the ways in which people use digital technologies to communicate about, and situate themselves in relation to, aspects of human and social experience such as gender and sexuality (Elm 2007; Halonen and Lepänen 2017; Milani 2013), political discourse (Chiluwa 2012; Mayo and Taboada 2017; Ross and Rivers 2017) and health communication (Brookes and Baker 2017; Harvey et al. 2013; Mullany et al. 2015).

This chapter will consider what English language studies can reveal about the human (and especially the social) world through a focus on *digital interactions*, defined here as multi-participant conversations and exchanges that make use of a computer, mobile phone or other electronic device. It will show that digital spaces in which individuals come together to interact, share information and form social bonds, such as social media platforms, discussion forums and blogging sites, can offer unprecedented access to multiple voices and competing perspectives in relation to a range of human and social concerns. As a result, they can provide fruitful ground for researchers to develop a deeper understanding of current social norms and trends, new and emerging communicative practices and the development and behaviour of social groups.

The chapter begins with a brief summary of recent English language research that has critically examined social issues such as discrimination and conflict as they are played out and negotiated in and through digital interactions. It will then introduce a theme that has been relatively little explored at the meeting point of English language and the digital humanities: *parenthood*. My own study of Mumsnet Talk (www.mumsnet.com/Talk), the discussion forum of a popular UK parenting website, will shed light on this area, contributing to knowledge in the digital humanities about what it means to be a parent and a mother in a digital age. My exploration of this study begins with a discussion of the ethical issues

involved in researching a digital site, which is a key area of debate in the digital humanities (Burdick et al. 2012). The chapter will also briefly outline the qualitative methods that are used to explore Mumsnet Talk interactions, highlighting the relevance of qualitative research in the digital humanities. It will show how digital technologies have not only facilitated the interactions that form the basis of the Mumsnet study but have also supported the qualitative research process through the use of specialised software for qualitative data analysis. Finally, some of the findings from the study will be illustrated through an exploration of the linguistic resources Mumsnet Talk contributors use to position themselves and others as individuals and as parents, and a consideration of how their involvement in the forum affects their access to these resources.

Current contributions and research

Research exploring the language of digital interactions in a range of contexts has revealed a great deal about the forms such communication can take in terms of specific linguistic devices, structures and styles. Lee (2007) and Nishimura (2007) have documented some of the linguistic and stylistic features of email and instant messages in Hong Kong and Japanese bulletin board messages, respectively, showing how writers use a range of creative strategies to communicate effectively in a digital medium. For example, Lee (2007: 191) has shown that, in emails and instant messages, Hong Kong users have represented Cantonese through innovative devices such as literal translations and romanised Cantonese, as well as traditional Cantonese and Chinese characters. Danet et al. (1997), Deumert (2014) and Katsuno and Yano (2007) have explored the potential for creativity, playfulness and expressivity in a range of social networking sites and discussion forums based in the United States, South Africa and Japan. For example, Danet et al. (1997: np) explore a sequence of early Internet Relay Chat in which playful effects are created through the use of basic keyboard functions, such as the repetition of the letter s in ‘ssssssssssss’ and the use of asterisks, as in ‘*puff* *hold*’, to simulate the sounds and gestures involved in smoking marijuana. Katsuno and Yano (2007: 284) have pointed to the diversity of meanings that can be generated with *kaomoji*, the Japanese equivalent of emoticons, which express complex expressions, gestures and emotions by creating pictures using basic keyboard functions, such as ‘\(@o@)/’ to represent an expression of ‘sheer fright, eyes wide in astonishment, hands raised’.

In recent years, linguists have tended to be concerned not only with the micro-linguistic features and styles *of* digital interactions but also with the macro-social and political issues, group practices, communities and identities that are explored and created *through* digital interactions. This kind of user-focused sociocultural approach is perhaps demonstrated most clearly through research that explores both the creation and resistance of discriminatory practices in interactive digital contexts. For example, Chilwa (2012), Fozdar and Pedersen (2013) and Törnberg and Törnberg (2016) have examined discussions in customised websites, interactive blogs and forums to show how different individuals and groups have taken up both discriminatory and anti-discriminatory discourses online. Fozdar and Pedersen’s (2013) analysis of an Australian interactive blog about asylum seekers, for instance, suggests that online spaces where people can engage in public debate may be productive sites for the resistance of dominant norms and the negotiation of marginalised, emergent or counter-hegemonic discourses such as ‘bystander anti-racism’. Exploring the ways in which people negotiate competing norms and perspectives online is a recurring theme in studies of

language and digital interaction. For example, Gong (2016), Hall et al. (2012) and Lehtonen (2017) have examined some of the ways in which contributors to message boards and discussion forums have worked to negotiate and potentially transgress restrictive gender norms and expectations in their everyday digital interactions. However, many of these researchers are cautious about such findings, with Hall et al. (2012) and Gong (2016), for instance, suggesting that discourses of hegemonic masculinity and sexism continue to influence the negotiation of supposedly ‘new’, ‘alternative’ or ‘queer’ forms of masculinity in the online interactions they analyse.

Online communities and networks for parents, especially mothers, have been a source of particular interest for researchers in the social sciences and digital humanities who are concerned with group practices, communities and wider social issues around gender and parenthood. Much of this research has suggested that parenting forums around the world, such as Familjeliv in Sweden (Hanell and Salö 2017), Momstown in Canada (Mulcahy et al. 2015) and Happy Land in Hong Kong (Chan 2008), can offer safe spaces in which women can explore motherhood on their own terms, express their personal perspectives, negotiate shared experiences with other mothers and even challenge dominant forms of knowledge around femininity, pregnancy and motherhood. However, a number of scholars have also pointed to the persistence of traditional and dominant norms and expectations within such digital contexts, suggesting that they do not facilitate such egalitarian and transgressive interactions as it may first appear. Madge and O’Connor (2006), for example, have pointed to the normative demographics of sites such as babyworld, the UK’s first parenting website (no longer in operation), where participants persistently introduce themselves as the main carer in a two-person heterosexual relationship. Boon and Pentney (2015) have shown that contributors to the U.S. site BabyCenter (www.babycenter.com) present themselves in similar ways, pointing to the exclusive portrayal of white, cisgender, heteronormative families in users’ breastfeeding selfies. Such hegemonic norms of gender, sexuality and race, they suggest, are propagated by the architecture of the site itself, which portrays the same groups in its fixed, edited and advertising images.

Mumsnet Talk, the discussion forum of the popular British parenting website Mumsnet, has been identified as a particularly relevant site for the study of contemporary motherhood in a British context. Pedersen and Smithson (2013) and Pedersen (2016) have suggested that this forum, like many others around the world, provides a space in which women can shape their identities as mothers and resist dominant ideals and expectations of femininity and motherhood. Pedersen’s (2016) analysis of 50 Mumsnet Talk threads that use the term ‘good mother’ in the thread title, for example, shows that contributors construct a complex picture of ‘good motherhood’, including decisions that might not otherwise be considered in the best interests of the child, such as returning to work or giving up breastfeeding. Similarly, in her analysis of Mumsnet Talk threads about postnatal depression, Jaworska (2018) shows that members use a range of communicative resources, such as confessions and exemplums, to challenge and rework ideals of ‘good motherhood’. My own study of Mumsnet Talk (Mackenzie 2017a, 2018, 2019) can support the view that this site provides a forum for challenging and shifting norms of gender and parenthood, but with some important caveats, as I will show later.

In the following section, as part of a discussion of the ethics of Internet research in the digital humanities, I explore some of the practical, legal and ethical issues around using data from the Mumsnet Talk forum before presenting the overall approach taken in this study and some of its key findings.

Critical issues

Conducting ethical research in the digital humanities

Critical issues around the use of online data for research purposes have been the subject of much discussion by English language and digital humanities researchers, with particular attention being paid to the legality of using digital data and copyright rules (Pihlaja 2017); the public/private dichotomy and its implications (Giaxoglou 2017); issues of consent, including whether and how to ask for it (Rüdiger and Dayter 2017); and the use of data from specific sites such as Twitter (Ahmed et al. 2017), Tinder (Condie et al. 2017) and Mumsnet (Mackenzie 2017b). These reflections and discussions around the ethics of online research methods have in common a case-based, context-sensitive approach that takes into account the norms and regulations of the research site, the nature of its users and their interactions and the ways in which data are both collected and analysed, as well as practical considerations, including whether it is feasible to contact research subjects.

The ethical decision-making process for the Mumsnet study is based on such a context-sensitive approach. My self-reflexive observations of this forum, conducted as part of a qualitative, semi-ethnographic approach, led me to recognise certain norms of interaction and sharing within this context that subsequently affected my decisions about data sampling, informed consent and anonymity. For example, I realised that although this was a publicly accessible forum, contributors tended to address quite a specific audience and that most Mumsnet users would not expect a researcher to take an interest in their contributions. In addition, it became clear to me during the course of my observations that Mumsnet users often valued their sense of privacy and anonymity very highly, with many exercising their autonomy and agency in imaginative ways to control and shape the accessibility of their posts and the degree to which they were identifiable as single users. One of the most important decisions I made as a result of these considerations was to contact all of the Mumsnet users whose words I wished to quote and/or analyse in detail and ask for their informed consent.

The process of seeking consent from participants in the Mumsnet study involved careful planning. I initially contacted Mumsnet users via the site's private messaging system, a method chosen in collaboration with Mumsnet staff. I sent request messages to potential participants in batches of ten, spaced at 24-hour intervals, so that I could gauge participants' reactions before contacting the next group. Choosing the right words for this message involved several adjustments in response to opinions and concerns that participants raised as they responded. In addition, I made it very clear in this message that silence would *not* be interpreted as consent; in other words, contributors who did not respond were not included in the study. My message also gave contributors the option to have their usernames anonymised. Several participants chose to do so and were subsequently given pseudonyms that retained the spirit of their original username. Further explanation of this process and detailed insights about how I negotiated issues of privacy and anonymity are offered in Mackenzie (2017b) and Mackenzie (2019).

Communicating via private message was ultimately a very effective way of reaching out to Mumsnet users, who were usually quick to respond. There were also other, unanticipated benefits of contacting participants in this way. For example, some were keen to engage in conversation about my research. These informal discussions provided invaluable insights into my participants' opinions and feelings about Mumsnet Talk and about their interactions being used in academic research. These insights further informed my developing

understanding of the interactional norms within this forum. They also gave participants an opportunity to voice any concerns and ask questions – to have some sense that they were involved in the research process as participants rather than passive research subjects.

The following section offers further details about the Mumsnet study itself, including the nature of the research site and the research design, before presenting a sample analysis of one thread posted to Mumsnet Talk in June 2014.

The Mumsnet study

The parenting website Mumsnet was founded in the year 2000 by the British entrepreneur Justine Roberts, with the aim of ‘mak[ing] parents’ lives easier by pooling knowledge, advice and support’ (Mumsnet Limited 2019) and has since been steadily growing in popularity and status. The site hosts over 6 million visitors each month (Pedersen 2016), and its endorsement is highly sought after in commercial and political arenas. Mumsnet’s ‘family-friendly’ awards for companies ‘making life easier for parents’ in the UK and ‘Mumsnet best’ awards for products with high-ranking reviews (Mumsnet Limited 2019) are a testament to the site’s authoritative position on who, and what, can best meet the needs of UK families. The site has hosted a number of online discussions with politicians, and *The Times* even described the 2010 election as the ‘Mumsnet election’, due to its perceived influence amongst mothers as a key voting group (Pedersen and Smithson 2013). The Mumsnet website is divided into a number of different sections, but it is arguably best known for its discussion forum, Talk, an interactive space in which users can share details of their daily lives, make pleas and offers of support and exchange ideas or information.

Although the Mumsnet Talk forum has a relatively large number of users, the range of perspectives that are offered in this space is somewhat limited. Mumsnet claims to be a site for ‘parents’ in general (their tagline reads ‘by parents for parents’), but it targets users who identify themselves as female parents: as *mothers*. The name of the site, for example, employs the gendered category ‘mum’, and its logo seems to depict three women in ‘battle’ poses, armed with children or feeding equipment (see Figure 4.1). In addition, although Mumsnet is accessible around the world, it is very much a British site; its headquarters are in London, it is written exclusively in English and, in terms of topical discussion, it deals with many themes that are particular to a British context. Furthermore, demographic data collected by both Pedersen and Smithson (2013) and the 2009 Mumsnet census (see Pedersen and Smithson 2013) suggest that many Mumsnet users are working mothers with an above-average household income and a university degree. Any insights gained from the study of Mumsnet Talk are therefore not by any means generalisable to *all* mothers. However, the accessibility, popularity and influence of Mumsnet Talk means that discussions in this space are likely to influence wider norms and expectations around gender and parenting, and may even lead the way in terms of new concepts of ‘motherhood’. This makes it a particularly relevant context for examining the options available to women who are parents in contemporary British society.

The theoretical approach taken in this study of Mumsnet Talk is influenced by the work of Foucault (especially 1972, 1978), who emphasises the way powerful forces, often conceptualised as *discourses*, work to constitute the social world and the power relations, forms of knowledge and subject positions that come with it. My focus on language and post-structuralist theory distinguishes the Mumsnet study from other studies on parenting and motherhood online, such as those that have been introduced earlier. In short, this study aims to explore how Mumsnet users take up, negotiate and challenge discourses of gender,



Figure 4.1 Mumsnet logo

Source: Reproduced with the permission of Mumsnet Limited

parenting and motherhood in their digital interactions. It focuses on the complex interplay between multiple and competing forms of knowledge about gender and parenthood and the subject positions (or ‘ways of being an individual’, as Weedon (1997: 3) puts it) that are inscribed by these forms of knowledge. This focus is captured in the central research question ‘how do Mumsnet users position themselves, and how are they positioned, in relation to discourses of gender and parenthood in Mumsnet Talk interactions?’

The aims of the Mumsnet study are well served by qualitative methods. Until recently, research in the digital humanities has been dominated by ‘big data’ and quantitative research, but the value of ‘small data’ and qualitative methods is increasingly being recognised (boyd 2011; Evans and Rees 2012; Gelfgren 2016; Kitchin 2013). This chapter advocates for the development of qualitative methods in the digital humanities, showing that they are able to support rich, in-depth explorations of how social roles, norms and situations are constructed and negotiated in a digital age. In the case of the Mumsnet study, it is the norms and expectations around gender, parenting and motherhood, and the ways in which Mumsnet Talk users negotiate these norms through their digital interactions, that are under investigation. A qualitative approach is also aligned with the poststructuralist standpoint that is taken in this study, because it encourages and facilitates the researcher’s openness and flexibility, and is able to both acknowledge multiple voices in the data and embrace multiple interpretive possibilities.

A number of software packages can support qualitative data analysis (QDA) in English language and the digital humanities, such as ATLAS.ti, MAXQDA and NVivo. All of these tools can support the management and qualitative analysis of the multiple data types that are frequently found in digital contexts, including a range of text, visual and audio files. Within these programmes, the researcher creates a project space into which external data can be uploaded, coded, annotated, organised and searched. They all include visualisation tools, and some, such as MAXQDA, also offer a good range of functions for statistical analysis, making them particularly well suited to mixed-methods research. QSR

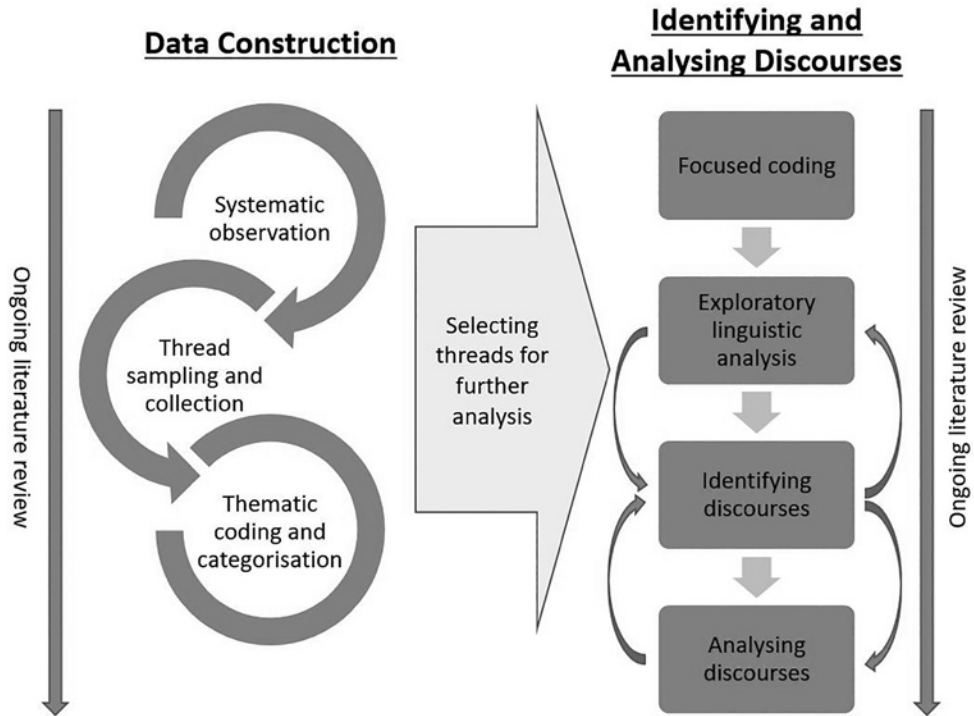


Figure 4.2 Research design for the Mumsnet study

International's NVivo 10 (2012) was identified as the software best equipped to serve the needs and exclusively qualitative research design of the Mumsnet study. Its NCapture function, which captures and preserves web pages in full, facilitated the collection of Mumsnet Talk discussion pages in a single file and in their original format, with text and images appearing exactly as they would to users of the forum. The key functionality of NVivo is coding, whereby passages of data can be stored and organised in 'nodes', which function as digital holding points for coded passages of text. Nodes can either be free, operating as separate entities, or organised in hierarchical relationships, so that related nodes can be grouped together within larger categories. The qualitative methods that are employed in the Mumsnet Talk study are briefly outlined later, including details of how coding within NVivo supported the research process.

The Mumsnet study can be divided into two stages: 'data construction' and 'identifying and analysing discourses', which are summarised in visual form in Figure 4.2 (see Mackenzie 2019 for a detailed overview of the research process in full, including theory, methods and data). In keeping with the qualitative paradigm, these processes advanced semi-iteratively, rather than being employed in a pre-determined, linear fashion.

The first stage of the Mumsnet study, 'data construction', was influenced by the qualitative traditions of both ethnography and grounded theory and served to enrich my understanding of both the Mumsnet Talk discussion forum and the research process itself. The naming of this stage is drawn from constructivist grounded theory (Charmaz 2014), which acknowledges that both research and the 'data' on which it is based are not fixed or stable concepts or entities, but are *constructed* by the researcher (also see Mason 2002). The

methods utilised during the process of data construction include ‘systematic observation’ (Androutsopoulos 2008) of the forum over a period of five months, which facilitated emerging insights about the benefits Mumsnet users gain from being members of the site, as well as the nature of Mumsnet Talk discussions themselves. During the observation period, threads were collected through purposive sampling (Oliver 2006; Robson 2011), where the aims of the study guided the selection of threads for further analysis. This process led to the creation of a corpus of 50 threads totalling just under 220,000 words. These threads were stored, coded and categorised in digital form within NVivo. Memos were also written and stored in digital form throughout this stage, providing opportunities to examine my role as a ‘constructor’ of data and serving as a record of my thoughts, reflections and developing interpretations.

The process of data construction was a springboard for the second stage of the Mumsnet study: ‘identifying and analysing discourses’. At the mid-point between these stages, I selected two threads for further detailed qualitative analysis – ‘Can we have a child exchange?’ and ‘Your identity as a mother’. These threads included several points of convergence between nodes relating to *gender* and *parenthood*, providing opportunities for the exploration of multiple and competing perspectives on these themes. The process of identifying and analysing discourses in these threads began with what Charmaz (2014) calls ‘focused coding’. This form of coding involved theorising about larger structures at work in the threads by identifying ‘theoretical codes’ that were indicative of discursive formations because they captured particular forms of knowledge, power and subjectivity. Some of the theoretical codes I identified from my analysis of ‘Your identity as a mother’ at this point, such as ‘total motherhood’ (which contains 118 references), were deemed important because of their prevalence in the thread, while others were considered analytically significant for other reasons. For example, the ‘equality between parents’ node is of particular interest because it contrasts so sharply with ‘total motherhood’. The ‘references to men’ node (which contains 21 references) is notable for its infrequency and points to a conspicuous absence of men across the thread. Such absences or marginalised themes were considered significant because it is not just what is central or dominant, but what is peripheral or marginalised, that can point to the presence of discourses (Sunderland 2000).

Focused coding provided the skeleton for my subsequent linguistic analysis of Mumsnet Talk threads: a point from which I moved forward and added flesh to the analytical bones I had constructed. The identification of key theoretical codes brought me to the next step in this stage of my research design, where I moved away from the coding process and focused on qualitative, micro-linguistic analysis of selected threads. This close linguistic analysis was operationalised through attention to how particular forms of knowledge about parenting and motherhood were constituted and what subject positions were available to Mumsnet users in relation to these forms of knowledge. Both language and other digital and semiotic forms were treated as key resources by which individuals could be positioned, or position themselves, in relation to these forms of knowledge and subjectivity. In the section that follows, I detail some of this analysis, showing how contributors to the ‘Your identity as a mother’ thread are positioned, and position themselves, in relation to discourses of *gendered parenthood*. The underlying principle of this analysis – that individuals can be positioned in multiple or contradictory ways by drawing on a range of resources to discursively position themselves and others through interaction – is influenced by Davies and Harré’s (1990) *positioning theory*.

Sample analysis: gendered parenthood in ‘Your identity as a mother’

Close analysis of the ‘Your identity as a mother’ thread led me to identify six discourses that are taken up and negotiated by its contributors (see Mackenzie 2019 for a detailed overview of these and other discourses identified in the Mumsnet study). These discourses are named as follows:

- 1 Gendered parenthood
- 2 Child-centric motherhood
- 3 Mother as main parent
- 4 Absent fathers
- 5 Equal parenting
- 6 Individuality

The discourse of ‘gendered parenthood’ is identified here as an ‘overarching’ discourse (Sunderland 2000) that takes in the discourses of ‘child-centric motherhood’, ‘mother as main parent’ and ‘absent fathers’. My analysis suggests that these discourses of gendered parenthood all work, in slightly different ways, to fix parents in binary gendered subject positions, as ‘mothers’ and ‘fathers’, and to restrict their access to a range of different subject positions. An example of this can be seen in the following posts, where contributors’ use of intensifiers suggests that their position as ‘mothers’ has near-total influence over their sense of self and is the only subject position available to them:

Post 3. *Cakesonatrain*. I think I am almost entirely Mum.

Post 11. *EggNChips*. As soon as I became a mum, I was 100% mum and loved it . . .

The discourses that remain – ‘equal parenting’ and ‘individuality’ – are arguably not gendered. However, gender is often foregrounded in the constitution of these discourses in the ‘Your identity as a mother’ thread. The competing relation between gendered and non-gendered discourses of parenthood can be seen in the following extracts, where non-gendered subject positions such as ‘parent’ and ‘person’ are taken up in opposition to the gendered subject position ‘mother’:

Post 15. *NotCitrus*. I feel much more of a parent than a “mum”

Post 13. *Crazym*. Hate being identified as “mum” . . . I have a name!!!! I am a person!!

In this section, I show how the gendered ‘mother as main parent’ and ‘absent fathers’ discourses are taken up in the ‘Your identity as a mother’ thread and how both the context of this thread and the discussion forum as a whole can make it difficult for contributors to position themselves outside of these discourses. I will then show how Mumsnet users take up a discourse that is oppositional and competing in this context: ‘equal parenting’. The analysis that is presented here focuses on participants’ use of categories and pronouns, two linguistic resources that emerged as particularly important for participants’ positioning of themselves and others in relation to discourses of gender and parenthood in this thread.

‘Mother as main parent’ and ‘absent fathers’

In the title and opening post of ‘Your identity as a mother’ (extract 1), pandarific invites contributors to address their self-positioning as ‘mothers’, exploring their ‘experiences’ of

motherhood, how motherhood ‘changes’ them and ‘their view of themselves’. She repeatedly uses the category ‘mum/mother’, which directly indexes gender, together with the second-person pronouns ‘you’ and ‘your’, explicitly positioning her audience as *female* parents. The title of the Mumsnet website itself, of course, deploys the same category; as noted earlier, this is a site that targets women, even though it is billed as a site by and for parents in general. By offering this category as the obvious, common-sense subject position, both the larger site more generally, and the ‘Your identity as a mother’ thread specifically, work to exclude parents who do not identify as mothers, such as male parents (and also carers who do not identify as parents), and limit the range of subject positions immediately available to their target demographic. The prevalence of the ‘mum’ category can be said to legitimise an overarching discourse of ‘gendered parenthood’, reinforcing a dichotomous gender divide between ‘male’ and ‘female’ parents.

Extract 1. Opening post to ‘Your identity as a mother’.

pandarific Sun 01-Jun-14 14:43:17

- 1 I’ve been reading a lot of fiction that deals with motherhood and family relationships
- 2 and I’m curious as to how it changes people, and their view of themselves. Has your
- 3 perception of who you are changed since you had children? How much of your identity
- 4 is bound up with being a mum? Do you think the strength of your desire to be a
- 5 mum/what stage in your life you had them affected the degree of the changes?
- 6 For some reason this has come out reading like an exam question – it’s not meant to
- 7 be! Just curious about people’s experiences.

Most contributors to this thread accept the gender-specific agenda set by pandarific. All participants present themselves as women, most identify as mothers and they tend to respond to the direct second-person address of the opening post with first-person singular pronouns such as ‘I’, ‘me’ and ‘my’. This pattern can be seen in posts such as cakesonatrains’ (extract 2), which almost exclusively employs first-person singular pronouns (see my emphasis in italics).

Extract 2. Post 3 to ‘Your identity as a mother’.

cakesonatrains Sun 01-Jun-14 15:07:15

- 1 *I think I am almost entirely Mum. My dc are still both under 3 so there’s a lot of*
- 2 *physical Mumming to do, with breastfeeding, nappies, carrying, bathing etc. I don’t*
- 3 *know if it will be less intense when they’re older, and I might let myself be a bit more*
- 4 *Me again, but right now I am almost refusing to have an identity beyond Mum.*
- 5 *I am a bit ‘old Me’ at work, but I’m part time now so there’s less of that too.*

Cakesonatrains use of the possessive pronoun ‘my’ in this post excludes any others from the ‘ownership’ of her children. She also defines the tasks of ‘breastfeeding, nappies, carrying, bathing etc.’, most of which (except for breastfeeding) can be carried out by any carer, using the gendered term ‘mumming’. Through her creative manipulation of the word ‘mum’, usually found in nominal form, she genders these parenting tasks, excluding a male parent (or indeed any other carer) from potential involvement. Her use of the term ‘breastfeeding’ – a task that is biologically gendered, rather than, for example, the gender-neutral ‘feeding’, together with her positioning of this gendered act at the beginning of a list of ‘mumming’ tasks, also works to exclude male or other non-breastfeeding carers by presenting these tasks as gender specific and integral to the role of ‘mum’.

The predominance of first-person singular pronouns in the ‘Your identity as a mother’ thread is not surprising, given pandarific’s exclusive emphasis on female parents in the title and opening post. But there are several recurring statements that could logically break away from this singular emphasis on contributors themselves and challenge the terms of the opening post. For example, within the clause ‘I have children’, which recurs in various forms across the thread (see the following examples), the pronoun ‘I’ could easily be replaced with ‘we’ or ‘our’ with little change to the content of the post, as it is in the excerpt from post 9.

Post 85. *I* am expecting *my* first (my emphases)

Post 9. *we* have only one child (my emphasis)

The excerpt from post 9 shows that one contributor to the thread *is* able to challenge the terms of the opening post while still offering a relevant response, through her use of plural reference to the self plus a second parent. Overall, however, broader patterns of category and pronoun use across the ‘Your identity as a mother’ thread, together with the more specific gendering of parental activities in posts like cakesonatrains’, work to create the impression that fathers are largely absent from contributors’ lives and that bringing up children is predominantly the domain of mothers, who have total responsibility for their children. These patterns point to the dominance of a pair of complementary discourses of gendered parenthood that restrict the range of subject positions available to both female and male parents: ‘mother as main parent’ and ‘absent fathers’.

It is important to note at this point that men and fathers are not completely absent from the ‘Your identity as a mother’ thread. Exploring moments at which men *are* made relevant, however, reveals that the ‘gendered parenthood’ and ‘mother as main parent’ discourses still dominate in many posts in which fathers are present, such as post 19 (see extract 3).

Extract 3. Post 19 to ‘Your identity as a mother’

MorningTimes Sun 01-Jun-14 20:29:39

- 1 I agree with cakes – I feel as if I am 99% ‘mum’ at the moment! I know it will pass
- 2 though, I am a SAHM [stay-at-home-mum] and am looking after three preschool DC [darling
- 3 children] all day (plus a fourth who is at school during the day) so I just don’t have the time or
- 4 the energy to be anyone else at the moment.
- 5 DH [darling husband] does a lot with the children too but at least he has a separate identity

- 6 because he is out at work and mixes with other people there.
- 7 I also struggle to spend time away from the DC for more than a night though. I have
- 8 friends who will happily go abroad on holiday without their DC but I just wouldn't want
- 9 to do that.

In this post, MorningTimes creates a clear physical separation between the main subject of each paragraph – ‘I’ (line 1), ‘DH’ (darling husband; line 5) and ‘I’ (line 7) – and uses exclusively singular pronouns when referring to herself or her husband. Through these linguistic and structural resources, MorningTimes positions herself and her husband as separate individuals. Further, her self-positioning in a gendered parental role, as a ‘mum’, through the relational clauses ‘I am 99% ‘mum’’ (line 1) and ‘I am a SAHM’ (line 2), contrasts with the positioning of her husband as someone who undertakes parental activities – who ‘does a lot with the children’ (line 5). The acronym ‘SAHM’ further positions her in the private sphere (‘at home’), whereas she positions her husband in an opposing ‘work’ space. The ‘absent fathers’/‘mother as main parent’ discourses do not operate in conjunction here, but it can be said that the ‘part-time father’/‘mother as main parent’ discourses (originally identified by Sunderland 2000) do: MorningTimes is the ‘main parent’ who devotes all of her time (‘all day’) and indeed herself (‘I am 99% ‘mum’’) to her children, while her husband is the ‘part-time parent’ who devotes much of his time elsewhere (‘he is out at work’) and whose status as a parent determines only what he *does*, not who he *is*.

‘Equal parenting’

The analysis that is presented earlier shows that male parents are largely absent from the discussions of ‘Your identity as a mother’ and that where they are made relevant, they are often positioned as oppositional to female parents, thus legitimising the ‘gendered parenthood’, ‘mother as main parent’ and ‘absent fathers’ discourses. However, there are some contributors to the thread who work to position male parents differently – as part of an equal, stable unit that acts jointly. This can be seen in extract 4, where EggNChips uses a wide range of linguistic resources, including inclusive pronouns such as ‘we’, ‘our’ and ‘us’, to present herself and her ‘DP’ (darling partner) as a unit with equal parental roles.

Extract 4. Post 11 to ‘Your identity as a mother’

EggNChips Sun 01-Jun-14 18:49:33

- 1 I was ready to become a mum when I had my DS [darling son], it was well worth waiting for –
- 2 we’d mellowed as a couple and both completed our post grad courses/worked up to a good
- 3 place in employment and by the time he arrived, everything felt right.
- 4 As soon as I became a mum, I was 100% mum and loved it; threw myself in to just that. Then
- 5 slowly over time, returning to work initially part time, then more or less full time, I’m more
- 6 “me”.

- 7 I'm of course a mum at home but DP [darling partner] does equal amounts of parenting and
 8 between us we allow each other to do our own things (so I play for a sports team, do stuff (*sic*)
 9 the NCT [National Childbirth Trust], and regularly organize a meal out with my girlfriends; he's
 10 training for a sport thing and also meets his friend about an ongoing project). We also try and
 11 have a date night or some time on our own once in a while. DS has changed us, but only
 12 priorities, rather than us as people.
 13 Now DS is 2.8, I'm 50% mum and 50% me, I love my job, love my friends, love Dp, and my
 14 sports and there is so much more to me than being a parent.

Although EggNChips begins her post by focusing on herself and whether motherhood has changed her ('I was ready to become *a mum* when I had *my* DS'), she introduces her partner through the use of the plural pronoun 'we' in line 2. This pattern continues throughout EggNChips's post. For example, in the third paragraph (lines 7–12), she begins by focusing on herself 'as a mum' and using the first-person pronoun 'I' (line 7). As before, however, she follows this reference to self with a reference to herself and her DP as a unit, through the use of inclusive pronouns ('us', 'we', 'each other', 'our', 'people') and a preposition that situates them in close relation to one another ('between'). EggNChips goes on to list the 'things' that they allow each other to do, and in doing so presents them each as separate actors: 'I' and 'he' (lines 8–10). However, her use of lexically and syntactically similar sentences to list these activities, as in the clauses 'I play for a sports team' (line 8) and 'he's training for a sport thing' (lines 9–10), again emphasises the equality between herself and her partner. These constructions imply that they take part in an equal range of activities outside of the home and that their interests are very similar, therefore again positioning them as *equal parents*. Her use of the word 'equal' (line 7) to qualify the 'amounts of parenting' they each do makes this positioning explicit. By making her 'DP' relevant to her response, EggNChips suggests that 'having children' is not a one-way relationship between her and her child: it has also involved her partner. In this way, she positions them both as subjects of a discourse of 'equal parenting', which competes with 'mother as main parent' and 'absent fathers' by offering female and male parental subject positions on equal terms.

Despite EggNChips's claims to parental equality, however, her positioning of her partner is in many ways *not* equal and in some ways works to reinforce the legitimacy of both the 'gendered parenthood' and 'mother as main parent' discourses. For example, the pre-clause qualifier 'of course' (line 7) implies that her position as 'mum' in the home environment is obvious and taken for granted. Further, in the statement 'DP does equal amounts of parenting' (line 7), EggNChips subtly positions herself as the main parent whose contribution is automatically assumed; the standard by which her partner's parenting contribution is compared. In addition, EggNChips persistently positions herself within the overarching discourse of 'gendered parenthood', as a 'mum', in the first line of every paragraph in her post. Like MorningTimes, however, she does not position her partner directly as a father or as a parent; he is described as someone who 'does . . . parenting' (line 7).

The sample analysis that is presented here shows how users of Mumsnet Talk take up, negotiate and challenge the dominant 'mother as main parent' and 'absent fathers' discourses in their digital interactions. In response to the central research question 'how do Mumsnet users position themselves, and how are they positioned, in relation to discourses of gender

and parenthood in Mumsnet Talk interactions?’ I have begun to show that, despite moments of resistance and challenge, discourses of gendered parenthood persistently work to position parents as gendered subjects who are either ‘main parents’, ‘absent fathers’ or ‘part-time fathers’. These discursive positions are often realised through referential devices such as categories and pronouns, which have acquired such common-sense legitimacy that they are difficult to avoid. The findings that are outlined in this section illustrate some of the demands and expectations that parents experience, showing why, for example, it continues to be difficult for male parents to be positioned as primary carers who make significant sacrifices for their children.

Future directions

This chapter has shown that digital interactions provide fruitful data sources for English language and digital humanities research that explores pressing issues for society, such as the question of whether mothers can access new and potentially more egalitarian ways of being a parent through their online discussions. As noted at the start of this chapter, Mumsnet members represent a relatively limited demographic, meaning that insights gained from this study are by no means generalisable to all mothers or all contexts. The popularity and influence of the forum, however, together with the fact that similar sites exist around the world, suggests that these findings can contribute to better understanding of the ways in which parenthood can be expressed and explored in a digital age.

There are many important social groups and themes that remain under-explored through the study of Mumsnet that is presented here. For example, the experiences of marginalised groups who would otherwise be geographically and temporally disparate can be brought to the fore through the exploration of digital interactions, since digital spaces can offer key sites for such groups to access peer networks and community support. In relation to the theme of parenthood, future research may therefore consider how marginalised and alternative family configurations might be negotiated and new definitions of the ‘family’ carved out through digital interactions. In addition, developing technologies continue to offer new, emergent and shifting forms of interaction and a wide range of semiotic resources through which these kinds of negotiation may be played out, such as memes (Ross and Rivers 2017; Zappavigna 2012), images and selfies (Zappavigna 2016; Zappavigna and Zhao 2017) and sound and music (Machin and van Leeuwen 2016). The continual development of such varied and multiple semiotic modes opens up new ways of relating to others and to society, creating new avenues for future research that explores the significance of digital interactions in our lives.

Further reading

- 1 Bouvier, G. (ed.) (2016). *Discourse and social media*. London and New York: Routledge.
This interdisciplinary collection focuses squarely on methodology, offering a range of discourse-analytic approaches to the analysis of social media in relation to wider social issues.
- 2 Page, R., Barton, D., Unger, J.W. and Zappavigna, M. (2014). *Researching language and social media: A student guide*. London and New York: Routledge.
This guide is extremely useful for those who are relatively new to the exploration of language and social media (or other digital discourse). It focuses on practical advice and guidance and is interspersed with anecdotes and advice from prominent scholars in the field.
- 3 Georgakopoulou, A. and Spilioti, T. (eds.) (2015). *The Routledge handbook of language and digital communication*. London and New York: Routledge.
This handbook offers a comprehensive overview of recent theories, methods, contexts and debates in the language-focused study of digital communication.

References

- Ahmed, W., Bath, P. and Demartini, G. (2017). Using twitter as a data source: An overview of ethical, legal and methodological challenges. In K. Woodfield (ed.), *The ethics of online research*. Bingley: Emerald Publishing Ltd, pp. 79–107.
- Androutsopoulos, J. (2008). Potentials and limitations of discourse-centred online ethnography. *Language@Internet* 5: article 9.
- Boon, S. and Pentney, B. (2015). Virtual lactivism: Breastfeeding selfies and the performance of motherhood. *International Journal of Communication* 9: 1759–1774.
- boyd, d. (2011). Social network sites as networked publics: Affordances, dynamics, and implications. In Z. Papacharissi (ed.), *A networked self: Identity, community and culture on social network sites*. New York and London: Routledge, pp. 39–58.
- Brookes, G. and Baker, P. (2017). What does patient feedback reveal about the NHS? A mixed methods study of comments posted to the NHS Choices Online service. *BMJ Open* 7(4): 1–12.
- Burdick, A., Drucker, J., Lunenfeld, J., Presner, T. and Schnapp, P. (2012). *Digital humanities*. Cambridge, MA and London: The Massachusetts Institute of Technology Press.
- Chan, A.H. (2008). “Life in happy land”: Using virtual space and doing motherhood in Hong Kong. *Gender, Place & Culture* 15(2): 169–188.
- Charmaz, K. (2014). *Constructing grounded theory* (2nd ed.). Los Angeles, London, New Delhi, Singapore and Washington, DC: Sage Publications.
- Chiluwa, I. (2012). Social media networks and the discourse of resistance: A sociolinguistic CDA of Biafra online discourses. *Discourse & Society* 23(3): 217–244.
- Condie, J., Lean, G. and Wilcockson, B. (2017). The trouble with tinder: The ethical complexities of researching location-aware social discovery apps. In K. Woodfield (ed.), *The ethics of online research*. Bingley: Emerald Publishing Ltd, pp. 135–158.
- Danet, B., Ruedenberg-Wright, L. and Rosenbaum-Tamari, Y. (1997). “Hmmm . . . Where’s that smoke coming from?” writing, play and performance on internet relay chat. *Journal of Computer-Mediated Communication* 2(4).
- Davies, B. and Harré, R. (1990). Positioning: The discursive production of selves. *Journal for the Theory of Social Behaviour* 20(1): 43–46.
- Deumert, A. (2014). The performance of a ludic self on social network(ing) sites. In P. Seargeant and C. Tagg (eds.), *The language of social media: Identity and community on the internet*. Basingstoke: Palgrave MacMillan, pp. 23–45.
- Elm, M.S. (2007). Doing and undoing gender in a Swedish internet community. In M.S. Elm and J. Sundén (eds.), *Cyberfeminism in northern lights: Digital media and gender in a nordic context*. Newcastle: Cambridge Scholars Publishing, pp. 104–129.
- Evans, L. and Rees, S. (2012). An interpretation of digital humanities. In D. Berry (ed.), *Understanding digital humanities*. Basingstoke: Palgrave MacMillan, pp. 21–41.
- Foucault, M. (1972). *The archaeology of knowledge* (A. Sheridan Smith, trans.). London and New York: Routledge Classics.
- Foucault, M. (1978). *The history of sexuality volume 1: An introduction* (R. Hurley, ed.). London: Penguin.
- Fozdar, F. and Pedersen, A. (2013). Diablogging about asylum seekers: Building a counter-hegemonic discourse. *Discourse & Communication* 7(4): 371–388.
- Gelfgren, S. (2016). Reading Twitter: Combining qualitative and quantitative methods in the interpretation of Twitter material. In G. Griffin and M. Hayler (eds.), *Research methods for reading digital data in the digital humanities*. Edinburgh: Edinburgh University Press, pp. 93–110.
- Giaxoglou, K. (2017). Reflections on internet research ethics from language-focused research on web-based mourning: Revisiting the *private/public* distinction as a language ideology of differentiation. *Applied Linguistics Review* 8(2–3): 229–250.
- Gong, Y. (2016). Online discourse of masculinities in transnational football fandom: Chinese Arsenal fans’ talk around “gaofushuai” and “diaosi”. *Discourse & Society* 27(1): 20–37.

- Hall, M., Gough, B., Seymour-Smith, S. and Hansen, S. (2012). On-line constructions of metrosexuality and masculinities: A membership categorization analysis. *Gender and Language* 6(2): 379–404.
- Halonen, M. and Leppänen, S. (2017). “Pissis stories”: Performing transgressive and precarious girlhood in social media. In S. Leppänen, E. Westinen and S. Kytölä (eds.), *Social media discourse, (Dis)identifications and diversities*. New York and London: Routledge, pp. 39–61.
- Hanell, L. and Salö, L. (2017). Nine months of entextualizations: Discourse and knowledge in an online discussion forum thread for expectant parents. In C. Kerfoot and K. Hyltenstam (eds.), *Entangled discourses: South-North orders of visibility*. New York: Routledge, pp. 154–170.
- Harvey, K., Locher, M. and Mullany, L. (2013). “Can i be at risk of getting AIDS?”: A linguistic analysis of two Internet advice columns on sexual health. *Linguistik Online* 59(2).
- Jaworska, S. (2018). ‘Bad’ mums and the ‘untellable’: Narrative practices and agency in online stories about postnatal depression on Mumsnet. *Discourse, Context & Media* 25: 25–33.
- Katsuno, H. and Yano, C. (2007). Kaomaji and expressivity in a Japanese housewives’ chat room. In B. Danet and S.C. Herring (eds.), *The multilingual internet: Language, culture, and communication online*. New York: Oxford University Press, pp. 278–302.
- Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography* 3(3): 262–267.
- Lee, C.K.M. (2007). Linguistic features of email and ICQ instant messaging in Hong Kong. In B. Danet and S.C. Herring (eds.), *The multilingual internet: Language, culture, and communication online*. New York: Oxford University Press, pp. 184–208.
- Lehtonen, S. (2017). “I just don’t know what went wrong” – Neo-sincerity and doing gender and age otherwise in a discussion forum for Finnish fans of My Little Pony. In S. Leppänen, E. Westinen and S. Kytölä (eds.), *Social media discourse, (Dis)identifications and diversities*. New York and London: Routledge, pp. 287–309.
- Machin, D. and van Leeuwen, T. (2016). Sound, music and gender in mobile games. *Gender and Language* 10(3): 412–432.
- Mackenzie, J. (2017a). “Can we have a child exchange?” Constructing and subverting the “good mother” through play in Mumsnet Talk. *Discourse & Society* 28(3): 296–312.
- Mackenzie, J. (2017b). Identifying informational norms in Mumsnet talk: A reflexive-linguistic approach to internet research ethics. *Applied Linguistics Review* 8(2–3): 293–314.
- Mackenzie, J. (2018). “Good mums don’t, apparently, wear make-up”: Negotiating discourses of gendered parenthood in Mumsnet Talk. *Gender and Language* 12(1): 114–135.
- Mackenzie, J. (2019). *Language, gender and parenthood online: Negotiating motherhood in Mumsnet talk*. London and New York: Routledge.
- Madge, C. and O’Connor, H. (2006). Parenting gone wired: Empowerment of new mothers on the Internet? *Social & Cultural Geography* 7(2): 199–220.
- Mason, J. (2002). *Qualitative researching* (2nd ed.). London, Thousand Oaks and New Delhi: Sage Publications.
- Mayo, M. and Taboada, M. (2017). Evaluation in political discourse addressed to women: Appraisal analysis of Cosmopolitan’s online coverage of the 2014 US midterm elections. *Discourse, Context & Media* 18: 40–48.
- Milani, T.M. (2013). Are “queers” really “queer”? Language, identity and same-sex desire in a South African online community. *Discourse & Society* 24(5): 615–633.
- Mulcahy, C.C., Parry, D.C. and Glover, T.D. (2015). From mothering without a net to mothering on the net: The impact of an online social networking site on experiences of postpartum depression. *Journal of the Motherhood Initiative for Research and Community Involvement* 6(1): 92–106.
- Mullany, L., Harvey, K., Smith, C. and Adolphs, S. (2015). “Am I anorexic?” Weight, eating and discourses of the body in online adolescent communication. *Communication and Medicine* 12(1): 211–223.
- Mumsnet Limited. (2019). Available at: www.mumsnet.com, www.mumsnet.com/
- Nishimura, Y. (2007). Linguistic innovation and interactional features in Japanese BBS communication. In B. Danet and S. Herring (eds.), *The multilingual internet: Language, culture, and communication online*. Oxford and New York: Oxford University Press, pp. 163–183.

- Oliver, P. (2006). Purposive sampling. In V. Jupp (ed.), *The sage dictionary of social research methods*. London, Thousand Oaks and New Delhi: Sage Publications, pp. 244–245.
- Pedersen, S. (2016). The good, the bad and the “good enough” mother on the UK parenting forum Mumsnet. *Women’s Studies International Forum* 59: 32–38.
- Pedersen, S. and Smithson, J. (2013). Mothers with attitude – How the Mumsnet parenting forum offers space for new forms of femininity to emerge online. *Women’s Studies International Forum* 38: 97–106.
- Pihlaja, S. (2017). More than fifty shades of grey: Copyright on social network sites. *Applied Linguistics Review* 8(2–3): 293–314.
- QSR International Pty Limited. (2012). NVivo qualitative data analysis software. Available at: www.qsrinternational.com/nvivo/home.
- Robson, C. (2011). *Real world research: A resource for users of social research methods in applied settings* (3rd ed.). Chichester: John Wiley & Sons Ltd.
- Ross, A.S. and Rivers, D.J. (2017). Digital cultures of political participation: Internet memes and the discursive delegitimization of the 2016 U.S Presidential candidates. *Discourse, Context and Media* 16: 1–11.
- Rüdiger, S. and Dayter, D. (2017). The ethics of researching unlikeable subjects: Language in an online community. *Applied Linguistics Review* 8(2–3): 251–270.
- Sunderland, J. (2000). Baby entertainer, bumbling assistant and line manager: Discourses of fatherhood in parentcraft texts. *Discourse & Society* 11(2): 249–274.
- Törnberg, A. and Törnberg, P. (2016). Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum. *Discourse & Society* 27(4): 401–422.
- Weedon, C. (1997). *Feminist practice and poststructuralist theory* (2nd ed.). Malden, MA, Oxford and Victoria, Australia: Blackwell Publishing.
- Zappavigna, M. (2012). *Discourse of twitter and social media*. London: Continuum.
- Zappavigna, M. (2016). Social media photography: Construing subjectivity in Instagram images. *Visual Communication* 15: 271–292.
- Zappavigna, M. and Zhao, S. (2017). Selfies in “mommyblogging”: An emerging visual genre. *Discourse, Context & Media* 20: 239–247.

Multimodality I

Speech, prosody and gestures

Phoebe Lin and Yaoyao Chen

Introduction

In the age of the Internet, social media have become an important and popular outlet through which people express thoughts and share information about their everyday lives. The most popular social media website, YouTube, for example, records 300 hours of video uploads per minute globally. These videos cover a variety of topics from cooking and craft demonstrations to video diaries and live broadcasts. Twitter is another well-known example. Its 330 million active users generate, on average, 500 million tweets every day. These staggering figures highlight the importance of social media as valuable language data for understanding communication in modern life.

As computers, tablets, mobile devices and electronic communication continue to dominate modern life, researchers have been exploring ways in which today's computing technology and tools may be incorporated to benefit all aspects of humanities research. This thirst for digital humanities (DH) has brought a wealth of innovations and fundamental transformations in many humanities disciplines. In multimodality, for instance, DH has stimulated researchers to explore and exploit social media and other types of born-digital media (e.g. instant voice/video messaging, video conferencing, podcasts) as an abundant source of spontaneous multimodal research data. For decades, multimodality researchers have been facing problems arising from the lack, and the small size, of spontaneous multimodal data. With the advent of born-digital data, such problems can finally be overcome. In fact, so much born-digital audio-visual data have become available nowadays that multimodal researchers are no longer able to keep up with the pace at which new data are released. In terms of methods, multimodality researchers traditionally rely on manual effort to annotate, align and analyse data. In the tidal wave of DH, however, new software packages have been released to semi-automate many such multimodal data processing procedures. The availability of these software packages has provided favourable conditions for scaling up multimodality studies. This scaling-up of studies is vital for transforming theory and practice in the discipline of multimodality not only because researchers will have a greater diversity of data against which their theories about multimodal communication can be tested but also because the robustness of findings may increase with data size.

This chapter discusses the exciting new opportunities that today's computer technology and tools have brought to multimodality research, particularly to studies which examine patterns of speech prosody and gesture in spoken communication. We begin with a brief account of multimodality, followed by an introduction to current methods and frameworks for examining speech prosody and gesture as two major multimodal cues in spontaneous spoken communication. The chapter then identifies key challenges facing multimodality research with regards to data acquisition, processing and analysis. In the third section, we highlight some exciting projects which bring new computer technology and tools for overcoming these challenges. These new computer technologies and tools have been developed to enable the collection, annotation and interrogation of large-scale multimodal data collected in mundane speech contexts. In the 'Sample analysis' section, we demonstrate how these new intelligent tools can be used in combination in the multimodal analysis of a video blog posted on social media. Finally, we conclude with suggestions for further research. All the corpus resources mentioned in this chapter can be found after the 'Further reading' section.

Background

Research on multimodality

Multimodality is deeply rooted in social semiotics, systemic-functional linguistics and written discourse analysis (Gibbons 2012; Kress and Van Leeuwen 2002; Norris and Maier 2014; Page 2010). It differs fundamentally from traditional monomodal, word-centred linguistic research in that it not only investigates words in meaning-making but also how and why people in certain social contexts choose other modes to construct meaning (Jewitt et al. 2016). *Modes*, in the broadest sense, refer to the multitude of socially and culturally conditioned recourses for meaning-making and interaction (Van Leeuwen 2005). In spoken communication, for instance, modes that interlocutors may draw upon for making and extracting contextual meanings include texts, objects, space, colour, image, gesture, prosody, facial expression, eye gaze and body posture. The last decade has witnessed an increasing amount of attention to the multimodal discourse analysis of human interactions in different contexts (Bezemer 2008; Bezemer et al. 2014; Goodwin and LeBaron 2011; Hazel et al. 2014; Mondada 2014).

Traditionally, speech data are presented as textual transcripts saved as text or Word documents. Following data acquisition, the audio or visual recordings are transcribed into words. The textual transcripts are examined to generate new insights about how humans interact through speech. However, as researchers became increasingly aware that transcribed words alone are inadequate in capturing the contextual meanings in spoken communication (this is because human interactions are inherently multimodal), various types of annotations, such as interruptions and overlaps, pauses, pitch changes, gestures and particular body language, began to be added to the textual transcript to incorporate information about the use of other modes. Today, these multimodal transcripts are often saved as Extensible Markup Language (XML) files, which means that information about each mode can be presented as separate tiers in the transcript and, with the help of specialised software, interrogated independently or in conjunction with other tiers. In this chapter, we focus on speech prosody and gesture as two of the most well-researched modes of communication in multimodality.

An overview of speech prosody research

Speech prosody, commonly known as *tone of voice*, subsumes three supra-segmental aspects of speech, namely intonation, stress and rhythm. Acoustically, these three aspects correspond to and can be measured, respectively, in terms of changes in pitch, loudness and duration of a human speech sound wave (Cruttenden 1997). Prosody is vital when decoding spoken discourse because it carries a wide range of paralinguistic information that cannot be retrieved through examining words alone. These pieces of paralinguistic information encoded in prosody include the speaker's attitudinal meaning and self-identity, the utterance's intended function, syntactic structure and information structure and the discourse's organisational structure (see Wichmann 2015 for further discussion).

To illustrate the role of prosody in communicating attitudinal meanings, Crystal (2003) offers the example of nine ways of saying 'yes'. According to Crystal (2003), 'yes' with a full falling tone, for instance, signals emotional involvement, particularly excitement, surprise and irritation; the higher the onset of the tone, the more involved the speaker. 'Yes' with a mid-falling tone signals detachment and lack of commitment and excitement. 'Yes' articulated in a level tone signals boredom, sarcasm and irony (and so on). In terms of information structure, accentuation (i.e. the assignment of accent) often contributes to the differentiation between given and new information in an utterance. Generally speaking, lexical items representing new information tend to be accented; those representing old information tend to be deaccented. For example, in B's answer provided next, 'next week' constitutes new information and is thus accented. 'New York', on the other hand, constitutes given information and is deaccented.¹

A: When are you going to New York?

B: I'm going to New York NEXT WEEK.

Speech prosody has always been an important topic not only in phonetics and phonology but also in fields of applied research such as spoken discourse analysis, psycholinguistics and natural language processing. In spoken discourse analysis, researchers are particularly interested in speech prosody's roles in disambiguating attitudinal meaning and speech acts (e.g. Lin and Adolphs 2009), in addition to segmenting spontaneous speech into intonation units (Chafe 1987), spoken paragraphs (Wichmann 2000) and conversational turns (e.g. Couper-Kuhlen and Selting 1996; Szczepek Reed 2004). In psycholinguistics, speech prosody offers a unique window through which researchers may examine various aspects of language processing in the human brain, including lexical storage and retrieval in the mental lexicon (Lin 2010), stages of online speech planning (Warren 1999) and the capacity of a human's short-term memory (Chafe 1987). More recently, there is also a growing interest in the role of prosodic cues in driving first language acquisition (e.g. Fisher and Tokura 1996; Lin 2012). This new interest has given rise to studies investigating the prosodic properties of child-directed speech (e.g. Fisher and Tokura 1996; Shattuck-Hufnagel and Turk 1996), as well as usage-based models of language acquisition (Tomasello 2003). Finally, in natural language processing, researchers have been exploring means through which prosodic information can be incorporated to improve parsing accuracy (Bear and Price 1990), discourse segmentation (Levow 2004), speech summarisation (Jauhar et al. 2013) and the naturalness of computer-synthesised speech to the human ear (Pierrehumbert 1993; Rosenberg and Hirschberg 2009).

Approaches to speech prosody vary across fields. In phonetics and phonology, there are two fundamentally different systems for analysing, annotating and interpreting intonation patterns, namely the *auditory* and the *instrumental* approaches (Cruttenden 1997). Researchers following the auditory approach (Crystal 1969; Knowles et al. 1996) are trained to transcribe intonation by ear and to use their expert judgement to balance a range of factors in the transcription process, such as perceived pitch changes, individual differences in speech styles and the intended meanings of the speaker. Such an approach to the transcription and analysis of intonation has often been criticised for being unscientific and impressionistic. However, it remains the most frequently used and most effective method for annotating the prosodic features of multimodal spoken corpora. Corpora which were prosodically transcribed by ear often serve as data based on which automatic speech recognition and synthesis programmes were trained in natural language processing. Researchers following the instrumental approach to intonation (Gussenhoven 2004; Pierrehumbert 1980), on the other hand, emphasise precise and verifiable measurements in the transcription and analysis of intonation conducted with the help of tools such as Pratt (Boersma 2001). Following the instrumental approach, intonation is quantified as changes in fundamental frequency (F0), and the transcription indicates observable pitch fluctuations in the sound waves. Such an instrumental approach to intonation clearly has great potential applications for modern speech technology. However, the methods have so far only been applied to very narrow settings. Substantial work is needed to apply this approach reliably to the large-scale transcription of spoken discourse or corpora. There are already some new projects in this area (e.g. Rosenberg 2010; see also the ‘Current contributions and research’ section)

An overview of gesture studies

Gesture is a rather under-researched area in linguistics. However, there are numerous studies in psychology, cultural studies and anthropology investigating different types of gestures. For instance, some cultural and anthropological studies have fixated on highly conventionalised and cultural-specific gestures, which are usually categorised as emblems,² or quotable gestures (Brookes 2004, 2005; Kendon 1995). Each emblem can be regarded as a single ‘word’ with a discrete and cultural-specific meaning that can substitute the corresponding lexical item with a stable meaning across discourse contexts. An example is the well-known ring gesture with the thumb and index finger touching and the other fingers bended naturally. Its meanings differ across cultures (Desmond et al. 1979): it means ‘OK/good’ in many parts of Europe, but ‘zero’ in France. Another type of gesture is gesticulation (Kendon 2004), which refers to the improvising head, hand and arm movements performed alongside speech. The meanings of gesticulations vary depending on the discourse contexts (Calbris 2011). Compared to conventional emblems, it is believed that gesticulations tend to perform in a less conscious way and are frequently studied in psycholinguistics as a window to the mind (Duncan 2006; Kita et al. 2007).

Researchers in gesture studies have developed frameworks for analysing and describing gestures, which have been applied in the manual annotation of large multimodal corpora. The most commonly adopted framework for categorising and describing co-verbal gestures (also known as gesticulations) is developed by McNeill (1992, 2005).³ The framework classifies co-verbal gestures into iconic, metaphoric, deictic/pointing and beat gestures. Iconic gestures refer to motions performed to describe the features of the co-occurring speech parts, such as shapes of concrete objects or movements of actions. An example of an iconic gesture is a speaker outlining the shape of a book while saying the word. Metaphoric

gestures are motions which also resemble the core features in the co-occurring contents, but the contents are usually abstract (e.g. knowledge, concept, argument). Deictic gestures are motions executed with body parts (usually the fingers) pointing to an object, a location or a space. Beat gestures are up-down/left-right movements of the hand. Researchers (Kendon 2004; Kita et al. 1998; McNeill 2005) have also developed coding schemes for segmenting co-speech gesture phases based on the movements (e.g. preparation, stroke, retraction, pre-stroke hold, post-stroke hold). The stroke is the only compulsory phase, which requires relatively more effort in movement. It also aligns with the associated speech and embodies the meaning of the gesture. The part of speech that is co-expressive with the stroke phase is called the ‘lexical affiliate’ (Schegloff 1984), or the ‘idea unit’ (Kendon 2004). A further scheme worth mentioning is McNeill’s (2005: 274) illustration of different gesture spaces in front of the body (i.e. centre-centre, centre, periphery extreme-periphery). This framework has also been widely applied in gesture analysis and corpus annotation.

In linguistics, the studies of co-verbal gestures stem from a range of methodological and theoretical perspectives such as semantics (Calbris 2011), syntax and grammar (Fricke 2013; Harrison 2010; Steen and Turner 2013), pragmatics (Brookes 2005; Kendon 2004; Knight 2011), formulaic sequences (Chen 2015), metaphor (Chui 2011; Cienki and Müller 2008), and conversation analysis. (Goodwin and LeBaron 2011; Goodwin et al. 2002; Hazel et al. 2014; Mondada 2007). Although these studies do not draw on massive datasets, they are valuable for pushing the boundaries beyond word-centred investigation, expanding the existing knowledge and laying the foundations for future endeavours that pursue a stronger empirical value. One of the most well-known pieces of research is Kendon’s (1995, 2004) findings of the pragmatic functions of four gesture families in South Italy. Each gesture family shares the same core element in form and has stable functions across contexts. Among the four families, the Open Hand Supine family⁴ (or palm-up family with the meaning of offering, giving or receiving) and the Open Hand Prone family (or palm-down family with the meaning of negation, interruption, rejection) have been discovered and discussed in many other cultures.

A noticeable recent project is the exploration of form-meaning associations of many recurrent gestures (similar to Kendon’s gesture families) conducted by a group of German researchers working towards the linguistic potential of gestures (Fricke 2013; Müller et al. 2013). For example, based on more than 400 instances from a German corpus, Ladewig (2011) has built up a framework for describing the three functions of the Cyclic Gesture based on the discourse contexts, including delivering the meaning of continuity (referential function), word searching (cognitive function) and encouraging/requesting (pragmatic function). Chen and Adolphs (2017) have expanded and refined Ladewig’s research and conducted a multimodal corpus-based investigation of the circular gestures. Drawing on almost 600 circular gestures in an English corpus, they established the first lexical and grammatical framework based on the emerging co-gestural speech patterns. In view of research as such and a few other multimodal corpus-based projects (Chen 2015; Knight 2011; Kok 2017; Tsuchiya 2013; Zima 2014), we can witness a paradigm shift of combining both qualitative and quantitative approaches in linguistic gesture studies, aiming for increasing reliability and validity.

Critical issues and topics

As the body of research into patterns of prosody and gesture use continues to grow, multimodality researchers have been seeking ways to overcome various obstacles concerning

the data which have threatened the development of the discipline. These issues include, for example, the small size of multimodal data, the need for more types of spontaneous speech data and the tediousness of data alignment and annotation. These issues have hindered the scaling-up of multimodality research and limited the development and testing of new theories regarding multimodal communication. Therefore, there is urgency in tackling these issues.

Small data size

The small size and the lack of data have long been challenges in multimodality research. Most multimodality studies remain qualitative in nature and draw from a small number of speech extracts for evidence and examples. Studies reporting patterns of gesture use, for instance, typically draw on 50 or fewer instances as evidence.

In the past decades, many multimodal spoken corpora have been compiled as a significant aid of multimodal research, such as the *Nottingham Multimodal Corpus* (NMMC), the *AMI Meeting Corpus* and the *London-Lund Corpus*. These corpora are principled and balanced collections of audio-visual recordings designed to represent a wide variety of speech genres, including business transactions and meetings, radio broadcasts, public lectures, telephone calls and so on. The emergence of these corpora has facilitated the scaling-up of multimodality research and opened up new, quantitative approaches to multimodal communication.

As the number of multimodal corpora that can be used to investigate prosody and gesture continues to increase, the question then arises is the size. Even a multimodal corpus of a considerable size, such as the 7.5-million-word audio edition of the *British National Corpus*, might not be adequate, mainly for the reason that speech prosody and gestures are both complex phenomena subject to the effects of a myriad of individual, discourse, grammatical and lexical variables. Once the effects of extraneous variables are partialled out, very few samples would be left for further analyses. To illustrate this, consider Lin's (2013) study which examined the stress patterns of formulaic expressions in naturally occurring discourse. Using Wmatrix (Rayson 2003), 1580 expressions were identified in the 52,637-word Spoken English Corpus (SEC). However, after partialling out the effects of known extraneous variables (e.g. part of speech, tone unit structure), only 14 percent (i.e. 220 instances) were eligible for further analysis.

The need for large corpora is even more evident in the case of gesture studies. There are examples of large multimodal corpora, such as the NMMC (250,000 words) and the *AMI Meeting Corpus* (100-hour meeting recordings). However, the majority of multimodal corpora remain small with only a few videos. There is no available standard of how "big" a large multimodal corpus is; however, similar to the principle applied to the monomodal corpora, the more frequent the research target is, the smaller the multimodal corpora need to be (Adolphs 2008; Sinclair 2004). For frequently occurring gestures such as head nods, one hour can extract thousands of instances, and it will not require a large corpus to obtain enough evidence (Knight 2011). For less frequently occurring gestures such as the circular gestures (rotational movements of the hand), fewer than 1000 instances have been found in 10-hour videos (Chen and Adolphs 2017), and a large corpus will be necessary.

Generally speaking, a large multimodal corpus will be needed to conduct research on gesture. In addition to the need to consider the extraneous variables, some speakers just do not gesture frequently. The context and task of speaking can also affect the number of gestures. For instance, the increase of physical activities such as writing, drawing and typing may reduce talking and gesturing. In some supervision meetings in the NMMC, the

supervisors and students were discussing issues concerning data analysis and working quietly with software most of the time, which greatly reduced the rate of talking and gesturing. To generate a sufficient number of certain gestures, psycholinguists tend to design particular tasks (e.g. asking participants to retell cartoon clips that describe motion events) (McNeill 2005). The disadvantage of lab-based data is the lack of naturalness. Hence, decisions need to be made when balancing the quantity of the target instances and the authenticity of the research context.

The need for high-quality audio-visual recordings

There are many challenges to tackle in the process of multimodal data acquisition, one of which concerns the difficulty of collecting spontaneous speech data that comply with the audio-visual recording standards required for conducting multimodal analysis. For example, for accurate transcription and acoustic analyses of recorded sound waves, background noise needs to be kept to a minimum. To achieve this high standard of audio recording, each interlocutor needs a clip-on microphone. Furthermore, to record every detail of the interlocutors' non-verbal language (including head and hand gestures, facial expressions, eye gaze and posture), there needs to be a dedicated camcorder filming each participant. If video processing software will be used to aid the tracking of gestures, there will be additional requirements on lighting as well as the colour of the interlocutors' clothes. Satisfying all these technical demands is possible in a recording studio where participants are seated, but are rather challenging in mundane, everyday contexts, where people are moving around and the physical environment is much more complex.

Annotation takes ages

As multimodal corpus size increases, another issue that arises is the lack of computer tools for aligning and annotating audio-visual data streams automatically. Once multimodal speech data are acquired, the textual, audio and visual streams need to be aligned and annotated in preparation for systematic analyses. Currently, there is an array of tools for *displaying* time-aligned multimodal data streams in tiers (e.g. ELAN, Transana, DRS, UAM corpus tool). However, the processes of aligning the audio-visual streams with the textual transcript and annotating these streams are often done manually. Such manual annotation of multimodal features has the distinct advantage of offering a holistic, thorough and situated transcription of the multimodal features in the corpora. However, since the annotator needs to delve into all contextual cues in detail, the process of analysing a speech extract becomes a surprisingly tedious, painstaking and expensive process. In our experience, depending on the level of details required, it takes an hour to annotate the intonation and rhythm patterns in a single minute of speech. Similarly, depending on the density and complexity of the gestures examined, it can take a further hour to conduct a detailed annotation for one minute of a video. Apparently, without the aid of technology, the compilation of large multimodal corpora may be challenging due to the cost and manual labour.

Since the compilation and annotation of multimodal corpora are tedious and expensive, many multimodal studies are conducted using existing corpora. Some well-known and popular corpora for investigating prosodic patterns include the *London-Lund Corpus of Spoken English* (LLC), the audio edition of the *British National Corpus* (BNC), the *Intonational Variation in English* (IViE) corpus, the *Hong Kong Corpus of Spoken English* (HKCSE), the *Santa Barbara Corpus of Spoken American English* (SBCSAE), the *IBM/Lancaster SEC*,

the *Switchboard corpus*, the *CALLHOME corpus*, *CALLFRIEND corpus* and so on. Notable examples of multimodal corpora for investigating gesture patterns include the *AMI Meeting Corpus*, the NMMC, the *SmartKom corpus* and the *HuComTech corpus*.

Despite the lengthy list of examples provided here, the number of choices of multimodal corpora available to multimodal researchers remains limited. There are three reasons. First, the types and the levels of multimodal annotations available vary greatly by corpora. For example, the *AMI Meeting Corpus* annotates head and hand movements, the *SmartKom corpus* annotates head movements only and the *HuComTech corpus* annotates head shifts and handshape. As for corpora for research of prosody, the *LLC*, the *IViE* and the *HKCSE* annotate prosodic features; the audio edition of the BNC comes with audio recordings only; the *SBCSAE*, the *Switchboard corpus* and the *SEC* come with both audio recordings and annotations of prosodic features. Second, the annotation schemes adopted by multimodal corpora vary. For example, the transcribers of the *SEC* have developed their own assumptions and rules when annotating prosody; the transcribers of the *HKCSE* have adopted Brazil's (1978) Discourse Intonation system. Since the phonological assumptions of the corpora need to match those of a multimodal study, the number of multimodal corpora available to a researcher is indeed rather small. Third, not all multimodal corpora are publicly available. The NMMC, for instance, is not publicly available due to ethical reasons and/or requirements of the funding body (refer to Knight 2011 for further information regarding existing multimodal corpora).

Current contributions and research

While the previous section has illustrated the main challenges confronting multimodality research, the following discussion introduces some current projects that target these obstacles from the construction of larger multimodal corpora through to the innovative use of the latest computing technologies. They include projects that explore new sources of multimodal data, wearable devices for three-dimensional gesture capturing, tools for automatic alignment and annotation of multimodal data and software for displaying and searching between tiers of annotations. These achievements represent some of the most cutting-edge technologies in multimodality that are likely to inspire and equip other digital humanities researchers to go beyond textual analysis by drawing on 'born-digital' or real-life audio/video recordings of spoken language.

The Web as a new source of multimodal data

As mentioned earlier, compiling large multimodal corpora of high-quality recordings of everyday speech can be a challenge. As a result, many researchers have started to explore alternative sources of data which are 'born-digital'. Examples include instant voice and video messaging, telephone and video conferencing, social media and podcasts.

Lin (2016), for instance, proposed the use of social media such as YouTube videos as a new source of multimodal data and designed a computer tool that allows users to build and concordance their own multimodal corpus. The design of the tool builds upon the idea of the web-as-corpus, and exploits social media (such as video blogs) as an abundant resource of naturally occurring, user-generated media content. Another project that also exploits the potential of Internet media as a virtually limitless resource of born-digital spoken communication is Rooth et al. (2011) tool, which enables users to harvest, filter and prosodically annotate various types of audio recordings on the Web, including podcasts, radio broadcasts and other media-sharing sites.

Caution needs to be taken, however, when compiling multimodal corpora from born-digital data. For instance, as online videos are not necessarily tailored to gesture studies, the upper body and hand movements may only be partially visible, which turned out to be an issue for our sample analysis. Also, the recording environment (e.g. lighting, and the quality of the sounds and/or images) might be inadequate for fine-grained analysis using software. More importantly, many online videos have been rehearsed, performed and edited for better effect, making them unsuitable datasets for investigating spontaneous communication. Also, specific production and camera angles may impact upon our interpretation of the data. However, as the recording technologies continue to improve, born-digital materials may become an important source of data for speech, prosodic and gesture studies.

New wearable devices for collecting gesture data

Using the increasingly ubiquitous digital cameras and computers, sounds/prosody and images/gestures can now be recorded, replayed and captured as more solid evidence that is subject to further exploration and external review (Heath et al. 2010). In the meantime, the research contexts in which multimodal research is carried out have expanded from more instructed types of speech (e.g. speech, monologues in the lab) to more spontaneous forms of communication (e.g. real-life workplace communication) (Mondada 2013). To understand how people communicate in the real world, multimodal investigations of large datasets, or multimodal corpora containing many high-quality, systematically annotated videos collected from a range of communicative settings, are undoubtedly beneficial and essential.

Since gestures are performed in three-dimensional space, the attempt to capture gestures via any video recording device (e.g. camcorders, webcams) inevitably reduces the data to two dimensions and causes data loss (depth information), which restricts the accuracy of annotation and analysis. Videos are not machine readable, which results in the hard work of manual annotation. An alternative is the use of motion capturing systems (with/out reflective markers affixed to the targeted body parts) and data gloves, along with video recordings, to capture the precise position, posture, orientation and trajectory of the bodily movements (see Pfeiffer 2013a, 2013b for an overview of existing motion tracking technology and data gloves). The captured data are machine readable and thus reduce the manual work of annotation and analysis to a minimum. Projects using such technology include the D64 corpus, which captured gesture sequences and bodily movements in multi-party interaction (Campbell 2009) and the testing of head and hand moves in eliciting listener responses in multi-party discourse (Battersby and Healy 2010).

Earlier versions of these tools were bulky and difficult to set up, and as such had limited use in the laboratory setting. However, researchers have recently developed a wearable, light and inexpensive device called the *Wearable Acoustic Metre* (WAMs) (Seedhouse and Knight 2016). The device was designed to track and record sound and up-down/left-right motions of the hand by using tri-axial accelerometer in real time. The initial stage of test took place in a controlled setting, but the ultimate goal is to capture audio and motion data ‘in the wild’ in support of all types of verbal and non-verbal linguistic and interactive research. Real-life capture of multimodal behaviours is undeniably the most necessary technology in terms of assessing human interactions in distinctive communicative settings (Adolphs et al. 2011).

Automatic alignment and annotation tools

Whereas manual alignment and annotation of multimodal spoken corpora is still the standard practice, the situation is changing due to recent advances in forced alignment and automatic intonation annotation technologies.⁵ Once an audio-visual recording is transcribed, forced alignment tools such as *CMU Sphinx*, the *Penn Phonetics Lab Forced Aligner* (P2FA) and *Julius* can be applied to align automatically the textual transcript with the recordings. After that, the corpora can be annotated with the help of automatic intonation annotation tools such as *MoMel*, *ProsodyPro* and *AuToBI* (see sample workflow of multimodal data analysis in the ‘Sample analysis’ section). While there is always room for improvement, these tools are already very effective at reducing the time needed to prepare multimodal data for further analyses. More importantly, they provide a practical means through which the scale of multimodal data may be increased.

New tools for collating annotations and facilitating analyses

In addition to the tracking, recording and alignment tools mentioned earlier, numerous tools have long been available for displaying and annotating acoustic and visual streams of data. For instance, *ELAN* (Wittenburg et al. 2006) and *Anvil* (Kipp 2001) are widely used free software in gesture studies, especially from psycholinguistic and linguistic perspectives. This is because they offer the functions of marking multiple tiers or tracks of verbal and non-verbal data and frame-by-frame observations of motions.

Digital Replay System (DRS) (Knight and Tennent 2008) was an application developed and freely provided by the University of Nottingham that offered a unique function of concordancing verbal and non-verbal linguistic features that have been tagged beforehand. Therefore, it was relevant for the multimodal analysis of emerging behaviour patterns. DRS could be accompanied by a fieldwork tracker that could be downloaded on the phone in order to track GPS locations. Associated with the tracker could be any type of photographic, audio and video materials recorded by a user with time stamps and locations. However, DRS is no longer available. This leads to the issue of sustainability, one crucial factor that needs to be considered during the development of any software as the maintenance and updating of digital technology requires time, people and money.

Other tools with special features are also commercially available. These tools are products of world-class companies or institutions, and customers can benefit from long-term regular updates and technical support. Examples include *INTERACT* (particularly useful for coding online live data), *Observer XT* (for coding human and animal behaviour) and *Diver* (for panoramic video capture), among many others. With the abundance of tools, it is crucial for researchers to be aware of the theoretical and methodological assumptions underlying each tool in order to appropriately apply them to their own research.

Sample analysis

This section demonstrates one of the many ways in which the technological advances mentioned earlier may be applied to facilitating the compilation, annotation and analysis of prosody and gestures in born-digital datasets – in this case, video blogs (*vlogs*) from the social media website YouTube. The free tools applied in this sample analysis include Lin’s (2016) the *Web-as-a-Multimodal-Corpus Concordancing Tool*, the University of

Pennsylvania Linguistics Lab's *FAVE-align interface* (Rosenfelder et al. 2011), Rosenberg's (2010) *AuToBI*, Boersma's (2001) *Praat* and *ELAN* (version 4.9.4, Wittenburg et al. 2006).

To compile a multimodal corpus for sample analysis, we used Lin's (2016) *Web-as-a-Multimodal-Corpus Concordancing Tool*, which was designed to allow users to compile and concordance their own social media corpus. The expression *I think* is used as an example for demonstration due to its high frequency of occurrence in daily life. Based on the video addresses provided by the user, the tool builds the corpus and generates concordance lines with hyperlinks to the scenes in which the search term appeared. The corpus compiled here consists of 12 vlogs on YouTube, which we randomly sampled from the 3.28 million videos returned by searching for subtitled vlogs on the social media website. While each of the vlogs in the corpus can be examined in detail, we chose a vlog entitled 'Questions and Answers! [Tokyo Daily Vlog #25]' because it contains the highest frequency of *I think*, which in turn is the most frequent bigram in the corpus of English video blogs (see Figure 5.1 for the concordance of *I think* generated by the web-as-multimodal-corpus concordancing tool).⁶



Figure 5.1 A concordance of 'I think' generated using Lin's (2016) web-as-multimodal-corpus concordancing tool

As its title suggests, the vlog features a vlogger couple answering Facebook fans' questions about life as expatriates in Tokyo. In the 11-minute video, the couple took turns giving viewers their opinion on 15 questions ranging from Tokyo's transportation system and the couple's favourite foods. The high frequency of epistemic stance markers, such as *I think* (11 tokens), *I don't think* (3 tokens) and *I guess* (2 tokens), in the video also reflects the intended purpose of the communication.

The video was submitted to automatic prosodic annotation using Rosenberg's (2010) *AuToBI*, a tool which annotates pitch accents and phrase boundaries through analysing sound waves. The tool adopts an instrumental approach to intonation as its theoretical framework (see 'Background' section) and annotates using the Tones and Breaks Indices (aka *ToBI*, Pierrehumbert 1980). To run *AuToBI*, the audio channel of the vlog first needs to be aligned with the speech transcript at the word level. To conduct this process called *forced alignment*, we used *FAVE-align*, a web interface that implements the *Penn Phonetics Lab Forced Aligner (P2FA)* and takes the audio channel of the video and the transcript as input and generates an output file showing the alignment of the audio recording with the transcript at the levels of words and phones. This output .textgrid file from *FAVE-align* was then passed onto *AuToBI* for annotation in terms of intonational boundaries, pitch accents and break indices. The output from *AuToBI* is again a file with the .textgrid extension. All .textgrid files were then imported into *ELAN* together with the vlog video and the pitch measures extracted using *Praat*.

Figure 5.2 is a screenshot of the *ELAN* interface showing these various tiers of information time-aligned and ready to be interrogated. The tiers from top to bottom indicate (1) absolute pitch changes measured in hertz, (2) alignment at the word level for the male speaker S1, (3) alignment at the word level for the female speaker S2, (4) alignment at the

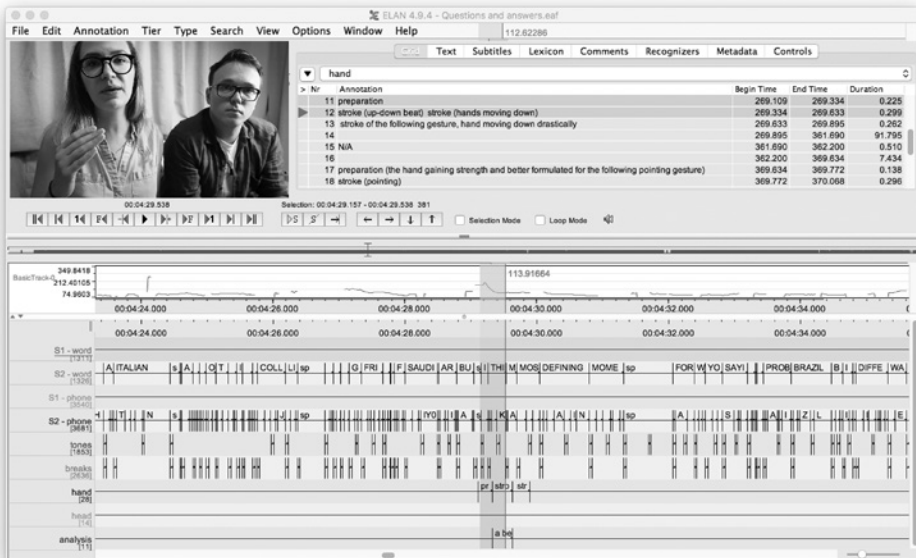


Figure 5.2 A screenshot of ELAN (version 4.9.4) showing tiers of information generated using an automatic tool

phone level for the male speaker S1, (5) alignment at the phone level for the female speaker S2, (6) annotation of tones by *AuToBI*, (7) annotation of breaks by *AuToBI*, (8) manual annotation of hand gestures, (9) manual annotation of head gestures and (10) researchers' comments.

Many linguistic questions in relation to speech, prosody and gesture can be addressed with a multimodal concordance as such. In this short sample analysis, we focus specifically on how prosody and gestures may offer insights into the use of the epistemic stance marker *I think* in context. As shown in the following concordance, *I think* was used 11 times in the vlog. (The code in brackets here indicate the speaker [male/female] and the corresponding time code measured in seconds.)

- 1 waste more money than buying shoes **I think** the preparation is really simple when you don't have to do anything specific you don't need (M62.193)
- 2 I've just been to like three or four **I think** now we're kind of stuck in Tokyo because of the company that we're running here (M174.570)
- 3 has really broadened my mind my horizons in the way that it's it's made me more international **I think** I realise how closed-minded I was with international issues (M240.252)
- 4 when I went to college like one of my good friends was from Saudi Arabia but **I think** my most defining moment for what you're saying was (F269.334)
- 5 my most defining moment for what you're saying was probably Brazil but in a different way **I think** it was more in the way of like realising that everything you believe isn't really (F275.603)
- 6 Is public transportation expensive in Japan **I think** I'm spending about a hundred dollars per month for transportation but then I'm not moving too much (M361.678)
- 7 but then I'm not moving too much around the city **I think** you're moving more yeah (M369.637)
- 8 tension with their neighbours that's that's the thing here **I think** for Polish people the best way to relate to that (M478.204)
- 9 an average Japanese person doesn't care about that **I think** American news sites Polish news sites love to have stories (M478.290)
- 10 **I think** the best time of year is fall it's very beautiful and it's more agreeable (M)
- 11 for me it's uncomfortable to be in this climate here **I think** you can still have fun you know it's not like don't ever come (M529.265)

As previous studies have suggested, *I think* can serve multiple functions (Aijmer 1997; Holmes 1990; Kärkkäinen 2003). It may signal a speaker's certainty (or uncertainty) about the truth value of a proposition, or used as a politeness strategy by softening the tone and building rapport. To disambiguate the contextual functions of the 11 instances of *I think* on the basis of the textual concordance lines alone would have been difficult. However, the task can be readily overcome when prosodic and gesture cues are considered. For instance, the function of *I think* in (1) seems ambiguous based on textual analysis. However, in view of the unusually high pitch of this instance and the stress on the pronoun *I* instead of *think*, it seems more certain that *I think* in (1) was articulated to express emphasis and the male speaker's confidence in his proposition.⁷ This interpretation of *I think* in (1) as a marker of certainty is also corroborated by our discourse analysis, which indicates that the male speaker was competing against the female speaker for the floor. In that immediate speech context, he foregrounded the certainty of his words and showed his intention to speak.

In addition to disambiguating the contextual meanings, prosodic cues provide important clues about the syntactic structure of naturally occurring discourse. In this vlog, speech prosody has provided vital clues about the attachment of the epistemic marker *I think*. Referring to (1), (2), (3), (5), (7) and (9) above, whether *I think* attaches to the preceding or following proposition is ambiguous. Upon listening to the audio channel and examining the prosodic annotations generated by *AuToBI*, however, it was clear that all instances of *I think* except (9) were positioned at the beginning of intonation units, which means that they were attached to the following propositions. This high tendency (over 90 percent) for *I think* to be used to begin, as opposed to end, propositions has also been observed in Kärkkäinen's (2003) study. When *I think* is attached to the end of a proposition and articulated in low pitch and low voice, as in the case of (9), it was probably used as a tone softener.

In terms of gestures, all the head and hand movements have been segmented manually in *ELAN* into five phases: preparation, pre-stroke hold, stroke, post-stroke hold and retraction (Kendon 2004; Kita et al. 1998; McNeill 1992, 2005). Among the five phases, the stroke phase is the only mandatory component for a gesture to exist, which constitutes, more or less, the most emphatic component in the entire gesture. The co-expressiveness between gesture and speech depends on the temporal coordination between the stroke phase and the corresponding speech part (e.g. a finger pointing upwards while saying 'up').

Perhaps due to the invisibility of the upper bodies of both speakers and the limited number of instances (12), only two instances were found co-occurring with two strokes: one is an up-down beat stroke in (4) (with the hand shape in Figure 5.2); the other up-down head-nodding movement in (10). Both strokes start and complete at the exact time periods when the two cases of *I think* are uttered, representing a close temporal relationship between the two modes. Functionally, both the beat gestures and the head movements have been found embodying the meaning of confirmation and emphasis (Calbris 2011; Kendon 2004; Knight 2011; McNeill 2005), which is iconic of the intended meaning of certainty expressed in *I think*. The results suggest a potential association between *I think* and the beating movements, co-expressing the sense of self-assurance. Therefore, similar to acoustic features, gestures can also help disambiguate the meaning of *I think*.

Future directions

This chapter has offered an overview of the key research avenues that have emerged at the interface between multimodality and digital humanities. We introduced some recent projects that seek to overcome the problems concerning the size and scope of existing multimodal spoken corpus data through the use of new computer tools and devices. In the 'Sample analysis' section, we demonstrated how these tools may be applied systematically to analyse large-scale multimodal spoken data. In the analysis, we also underlined the significant roles of advanced digital search, alignment and annotation tools in interrogating the functional relationships between the words, prosody and gesture.

The chapter proposed that the huge amount of born-digital spoken data from the Web and social media platforms could be a great source for compiling large-size multimodal corpora. This opens new research opportunities for linguists. For instance, it can assist research on the various aspects of spoken language such as vague language, spoken grammar, formulaic sequences, and metaphor, extending their analysis from texts to other modes of communication. It could further the investigations in pragmatics, that is, how pragmatic functions can be realised in spoken language through words and other modes such as prosody, gesture, facial

expressions and so on. It will also have positive repercussions for spoken discourse analysts and conversational analysts as, with multimodal corpora, the unfolding of spoken communication in different settings can be examined in videos with much richer information beyond words. Similarly, many important issues relating to sociolinguistics and critical discourse analysis (e.g. power, gender, race) can be explored drawing on multimodal spoken corpora in different contexts (e.g. how prosody, gesture and posture can mediate power relationships). Whereas traditional studies on spoken language rely on the same type of data, methods and frameworks used in written language, the availability of large multimodal spoken corpora may well lead to a conceptual, theoretical and methodological shift in most strands of research relevant to spoken language as researchers are trying to explore new approaches and frameworks to describe this new type of data (Adolphs 2008).

As we strive for multimodal research with greater depth and scale, new digital equipment and software are urgently needed to better track and record three-dimensional human behaviours and to realise (semi-)automatic alignment, mark-up and analysis. The achievement of this goal requires the collaboration between researchers from linguistics, computer science and other people in DH. The exciting projects introduced in this chapter are all important steps in this direction. These efforts in expanding data sources and analytical frameworks and computer tools will continue to grow and have a great impact on both the research community and the public.

Notes

- 1 Another key factor of accentuation in spontaneous speech is the distinction between broad and narrow focus. See Cruttenden (1997) for details.
- 2 Refer to McNeill (1992, 2005) for definitions of different types of bodily movements, i.e. gesticulation, emblems, pantomime and sign language.
- 3 In the 2005 version, McNeill defined them as gesture dimensions rather than categories as they are not exclusive to each other.
- 4 The Open Hand Supine family is also known as the Palm Up Open Hand gesture (PUOH) (McNeill 1992, 2005).
- 5 See <https://github.com/pettarin/forced-alignment-tools> for a list of forced alignment tools.
- 6 The url of the vlog is www.youtube.com/watch?v=fPIKWlet5FA
- 7 Previous studies, including Kärkkäinen (2003), have suggested that stress placement on *I* tends to signal the speaker's certainty, whereas stress placement on *think* tends to show the speaker's uncertainty.

Further reading

Web-as-a-Multimodal-Corpus Concordancing Tool

<https://www.idiomstube.com/corpus>

Forced alignment tools

- *CMU Sphinx* <http://cmusphinx.sourceforge.net/>
- *Penn Forced Aligner P2FA* www.ling.upenn.edu/phonetics/p2fa/
- *Julius* http://julius.osdn.jp/en_index.php

Automatic prosody annotation tools

- *MOMEL* www.ortolang.fr/market/tools/sldr000031/v3
- *ProsodyPro* www.homepages.ucl.ac.uk/~uclyyix/ProsodyPro/
- *AuToBI* <http://eniac.cs.qc.cuny.edu/andrew/autobi/>

Display and interrogation of collated multimodal data

- *Interact* www.mangold-international.com/en/software/interact
- *Observer XT* www.noldus.com/animal-behavior-research/products/the-observer-xt
- *Diver* <http://diver.stanford.edu/home.html>
- *Transana* www.transana.com
- *Digital Replay System (DRS)* <http://thedrs.sourceforge.net/>

References

- Adolphs, S. (2008). *Corpus and context : Investigating pragmatic functions in spoken discourse*. Amsterdam and Philadelphia: John Benjamins.
- Adolphs, S., Knight, D. and Carter, R. (2011). Capturing context for heterogeneous corpus analysis: Some first steps. *International Journal of Corpus Linguistics* 16(3): 305–324.
- Aijmer, K. (1997). *I think* – an English modal particle. In T. Swan and O.J. Westvik (eds.), *Modality in the Germanic languages: Historical and comparative perspectives*. Berlin: Mouton de Gruyter, pp. 1–47.
- Battersby, S.A. and Healy, P.G.T. (2010). Using head movement to detect listener responses during multi-party dialogue. In *Proceedings of the LREC workshop on multimodal corpora*. Malta: Mediterranean Conference Centre, pp. 11–15.
- Bear, J. and Price, P. (1990). *Prosody, syntax and parsing*. Paper presented at the Proceedings of the 28th annual meeting on Association for Computational Linguistics, Pittsburgh, Pennsylvania, 6–9 June.
- Bezemer, J. (2008). Displaying orientation in the classroom: Students' multimodal responses to teacher instructions. *Linguistics and Education* 19(2): 166–178.
- Bezemer, J., Cope, A., Kress, G. and Kneebone, R. (2014). Holding the scalpel: Achieving surgical care in a learning environment. *Journal of Contemporary Ethnography* 43(1): 38–63.
- Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glott International* 5(9/10): 341–345.
- Brazil, D. (1978). *Discourse Intonation II*. Birmingham: English Language Research, University of Birmingham.
- Brookes, H. (2004). A repertoire of South African quotable gestures. *Journal of Linguistic Anthropology* 14(2): 186–224.
- Brookes, H. (2005). What gestures do: Some communicative functions of quotable gestures in conversations among Black urban South Africans. *Journal of Pragmatics* 37(12): 2044–2085.
- Calbris, G. (2011). *Elements of meaning in gesture*. Amsterdam and Philadelphia: John Benjamins.
- Campbell, N. (2009). Tools and resources for visualising conversational-speech interaction. In M. Kipp, J.C. Martin, P. Paggio and D. Heylen (eds.), *Multimodal corpora: From models of natural interaction to systems and applications*. Heidelberg: Springer, pp. 176–188.
- Chafe, W.L. (1987). Cognitive constraints on information flow. In R.S. Tomlin (ed.), *Coherence and grounding in discourse*. Philadelphia: John Benjamins, pp. 21–51.
- Chen, Y. (2015). *Towards the multimodal units of meaning: A multimodal corpus-based approach*. A paper presented at the British Association of Applied Linguistics (BAAL) Annual Meeting, Aston University, Birmingham, 3–5 September.
- Chen, Y. and Adolphs, S. (2017). *Multimodal corpus-based investigations of the gesture-speech relationship*. A paper presented at the 9th International Corpus Linguistics Conference, University of Birmingham, Birmingham, 24–28 July.
- Chui, K. (2011). Conceptual metaphors in gesture. *Cognitive Linguistics* 22(3): 437–458.
- Cienki, A. and Müller, C. (eds.) (2008). *Metaphor and gesture*. Amsterdam and Philadelphia: John Benjamins.
- Couper-Kuhlen, E. and Selting, M. (eds.) (1996). *Prosody in conversation*. Cambridge: Cambridge University Press.
- Cruttenden, A. (1997). *Intonation* (2nd ed.). Cambridge: Cambridge University Press.

- Crystal, D. (1969). *Prosodic systems and intonation in English*. Cambridge: Cambridge University Press.
- Crystal, D. (2003). Prosody. In D. Crystal (ed.), *The Cambridge encyclopedia of the English language* (2nd ed.). Cambridge: Cambridge University Press, pp. 248–249.
- Desmond, M., Collet, P., Marsh, P. and O’Shaughnessy, M. (1979). *Gestures: Their origins and distribution*. London: Jonathan Cape.
- Duncan, S. (2006). Co-expressivity of speech and gesture: Manner of motion in Spanish, English, and Chinese. In *Proceedings of the 27th Berkeley linguistic society annual meeting*. Berkeley, CA: Berkeley University Press, pp. 353–370.
- Fisher, C. and Tokura, H. (1996). Prosody in speech to infants: Direct and indirect acoustic cues to syntactic structure. In J.L. Morgan and K. Demuth (eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Hillsdale, NJ: Lawrence Erlbaum, pp. 343–363.
- Fricke, E. (2013). Towards a unified grammar of gesture and speech: A multimodal approach. In C. Müller, A. Cienki, E. Fricke, S.H. Ladewig, D. McNeill and S. Teßendorf (eds.), *Body-Language-Communication: An international handbook on multimodality in human interaction (volume 1)*. Berlin: De Gruyter Mouton, pp. 733–754.
- Gibbons, A. (2012). *Multimodality, cognition, and experimental literature*. London: Routledge.
- Goodwin, C. and LeBaron, C. (2011). *Embodied interaction: Language and body in the material world*. Cambridge: Cambridge University Press.
- Goodwin, M.H., Goodwin, C. and Yaeger-Dror, M. (2002). Multi-modality in girl’s game disputes. *Journal of Pragmatics* 34(10–11): 1621–1649.
- Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge: Cambridge University Press.
- Harrison, S. (2010). Evidence for node and scope of negation in coverbal gesture. *Gesture* 10(1): 29–51.
- Hazel, S., Mortensen, K. and Rasmussen, G. (2014). Introduction: A body of resources – CA studies of social conduct. *Journal of Pragmatics* 65: 1–9.
- Heath, C., Hindmarsh, J. and Luff, P. (2010). *Video in qualitative research: Analysing social interaction in everyday life*. London: Sage Publications.
- Holmes, J. (1990). Hedges and boosters in women’s and men’s speech. *Language & Communication* 10(3): 185–205.
- Jauhar, S.K., Chen, Y.-N. and Metze, F. (2013). *Prosody-based unsupervised speech summarization with two-layer mutually reinforced random walk*. A paper presented at The International Joint Conference on Natural Language Processing, Nagoya, Japan, 14–18 October.
- Jewitt, C., Bezemer, J. and O’Halloran, K. (2016). *Introducing multimodality*. London: Routledge.
- Kärkkäinen, E. (2003). *Epistemic stance in English conversation: A description of its interactional functions, with a focus on I think*. Philadelphia: John Benjamins.
- Kendon, A. (1995). Gestures as illocutionary and discourse structure markers in Southern Italian conversation. *Journal of Pragmatics* 23(3): 247–279.
- Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.
- Kipp, M. (2001). Anvil-a generic annotation tool for multimodal dialogue. In *Eurospeech 2001, Seventh European Conference on Speech Communication and Technology*, Aalborg, Denmark, pp. 1367–1370. Available at: <https://pdfs.semanticscholar.org/cc3b/557cd80a9c36083769c4e370a4b10567c073.pdf>.
- Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R. and Ishizuka, T. (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. *Language and Cognitive Processes* 22(8): 1212–1236.
- Kita, S., van Gijn, I. and van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth and M. Fröhlich (eds.), *Gesture and sign language in human-computer interaction: International gesture workshop bielefeld, Germany, September 17–19, 1997 Proceedings. Lecture notes in artificial intelligence*. Berlin: Springer-Verlag, pp. 23–35.
- Knight, D. (2011). *Multimodality and active listenership: A corpus approach*. London: Bloomsbury.

- Knight, D. and Tennent, P. (2008). Introducing DRS (The Digital Replay System): A tool for the future of Corpus Linguistic research and analysis. In N. Calzolari (ed.), *Proceedings of the 6th language resources and evaluation conference (LREC), Palais des Congrès, Marrakech*. Morocco: European Language Resources Association, pp. 26–31.
- Knowles, G., Wichmann, A. and Alderson, P.R. (eds.) (1996). *Working with speech: Perspectives on research into the Lancaster/IBM spoken English corpus*. London: Longman.
- Kok, K. (2017). Functional and temporal relations between spoken and gestured components of language: A corpus-based inquiry. *International Journal of Corpus Linguistics* 22(3): 1–26.
- Kress, G. and Van Leeuwen, T. (2002). Colour as a semiotic mode: Notes for a grammar of colour. *Visual Communication* 1(3): 343–368.
- Ladewig, S.H. (2011). Putting the cyclic gesture on a cognitive basis. *CogniTextes* 6. Available at: <https://journals.openedition.org/cognitextes/406>.
- Levow, G-A. (2004). *Prosodic cues to discourse segment boundaries in human-computer dialogue*. A paper presented at the Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, Boston, MA., 30 April–1 May.
- Lin, P. (2010). The phonology of formulaic sequences: A review. In D. Wood (ed.), *Perspectives on formulaic language: Acquisition and communication*. London: Continuum, pp. 174–193.
- Lin, P. (2012). Sound evidence: The missing piece of the jigsaw in formulaic language research. *Applied Linguistics* 33(3): 342–347.
- Lin, P. (2013). The prosody of idiomatic expressions in the IBM/Lancaster Spoken English Corpus. *International Journal of Corpus Linguistics* 18(4): 561–588.
- Lin, P. (2016). *Internet social media as a multimodal corpus for profiling the prosodic patterns of formulaic speech*. A Paper presented at Joint Conference of the English Linguistics Society of Korea and the Korea Society of Language and Information, Kyung Hee University, Seoul, South Korea, 28 May.
- Lin, P. and Adolphs, S. (2009). Sound evidence: Phraseological units in spoken corpora. In A. Barfield and H. Gyllstad (eds.), *Researching collocations in another language: Multiple interpretations*. Basingstoke: Palgrave Macmillan, pp. 34–48.
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago and London: The University of Chicago Press.
- McNeill, D. (2005). *Gesture and thought*. Chicago and London: The University of Chicago Press.
- Mondada, L. (2007). Multimodal resources for turn-taking: Pointing and the emergence of possible next speakers. *Discourse Studies* 9(2): 194–225.
- Mondada, L. (2013). Embodied and spatial resources for turn-taking in institutional multi-party interactions: Participatory democracy debates. *Journal of Pragmatics* 46(1): 39–68.
- Mondada, L. (2014). The local constitution of multimodal resources for social interaction. *Journal of Pragmatics* 65: 137–156.
- Müller, C., Bressemer, J. and Ladewig, S.H. (2013). Towards a grammar of gestures: A form-based view. In C. Müller, A. Cienki, E. Fricke, S.H. Ladewig, D. McNeill and S. Teßendorf (eds.), *Body-Language-Communication: An international handbook on multimodality in human interaction (volume 1)*. Berlin: De Gruyter Mouton, pp. 707–733.
- Norris, S. and Maier, C.D. (eds.) (2014). *Interactions, images and texts: A reader in multimodality*. Berlin: Walter de Gruyter.
- Page, R. (ed.) (2010). *New perspectives on narrative and multimodality*. London: Routledge.
- Pfeiffer, T. (2013a). Documentation of gestures with data gloves. In C. Müller, A. Cienki, E. Fricke, S.H. Ladewig, D. McNeill and S. Teßendorf (eds.), *Body-Language-Communication: An international handbook on multimodality in human interaction (volume 1)*. Berlin: Mouton de Gruyter, pp. 868–879.
- Pfeiffer, T. (2013b). Documentation of gestures with motion capture. In C. Müller, A. Cienki, E. Fricke, S.H. Ladewig, D. McNeill and S. Teßendorf (eds.), *Body-Language-Communication: An international handbook on multimodality in human interaction (volume 1)*. Berlin: Mouton de Gruyter, pp. 857–868.

- Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.
- Pierrehumbert, J. (1993). Prosody, intonation, and speech technology. In M. Bates and R.M. Weischedel (eds.), *Challenges in natural language processing*. Cambridge: Cambridge University Press, pp. 257–280.
- Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Unpublished doctoral dissertation, Lancaster University, Lancaster.
- Rooth, M., Howell, J. and Wagner, M. (2011). *Harvesting speech datasets for linguistic research on the web white paper*. Available at: <http://diggingintodata.org/file/381/download?token=ilkLGqq2>.
- Rosenberg, A. and Hirschberg, J. (2009). *Detecting pitch accents at the word, syllable and vowel level*. A paper presented at the Proceedings of NAACL HLT 2009, Boulder, CO, June.
- Rosenberg, R. (2010). *AuToBI – a tool for automatic ToBI annotation*. Available at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.463.4856&rep=rep1&type=pdf>.
- Rosenfelder, I., Fruehwald, J., Evanini, K. and Yuan, J. (2011). *FAVE (Forced Alignment and Vowel Extraction) program suite*. Available at: <http://fave.ling.upenn.edu>.
- Schegloff, E.A. (1984). On some gestures' relation to talk. In J.M. Atkinson and J. Heritage (eds.), *Structures of social action*. Cambridge: Cambridge University Press, pp. 266–296.
- Seedhouse, P. and Knight, D. (2016). Applying digital sensor technology: A problem-solving approach. *Applied Linguistics* 37(1): 7–32.
- Shattuck-Hufnagel, S. and Turk, A.E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research* 25(2): 193–247.
- Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.
- Steen, F. and Turner, M.B. (2013). Multimodal construction grammar. In M. Borkent, B. Dancygier and J. Hinnell (eds.), *Language and the creative mind*. Stanford: CSLI Publications/Center for the Study of Language & Information. Available at: http://cogweb.ucla.edu/crp/Papers/SteenTurner_2013_MultimodalConstructions.pdf.
- Szczepek Reed, B. (2004). Turn-final intonation in English. In E. Couper-Kuhlen and C.E. Ford (eds.), *Sound patterns in interaction*. Amsterdam: John Benjamins, pp. 97–119.
- Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
- Tsuchiya, K. (2013). *Listenership behaviours in intercultural encounters: A time-aligned multimodal corpus analysis*. Amsterdam: John Benjamins.
- Van Leeuwen, T. (2005). *Introducing social semiotics*. London: Routledge.
- Warren, P. (1999). Prosody and language processing. In S. Garrod and M.J. Pickering (eds.), *Language processing*. Hove: Psychology Press, pp. 155–188.
- Wichmann, A. (2000). *Intonation in text and discourse*. Harlow: Longman.
- Wichmann, A. (2015). Functions of intonation in discourse. In M. Reed and J. Levis (eds.), *The handbook of English pronunciation*. Malden, MA: Blackwell, pp. 175–189.
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. and Sloetjes, H. (2006). Elan: A professional framework for multimodality research. In *Proceedings of LREC 2006, fifth international conference on language resources and evaluation*. Genoa, Italy: Magazzini del Cotone Conference Center, 24–26 May. Available at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.136.5130&rep=rep1&type=pdf>.
- Zima, E. (2014). *English multimodal motion constructions: A construction grammar perspective*. Available at: <http://uahost.uantwerpen.be/linguist/SBKL/sbkl2013/Zim2013.pdf>

Multimodality II

Text and image

Sofia Malamatidou

Introduction

Multimodality, which can be defined as the combination of two or more semiotic resources (including language) within a particular communication event (O'Halloran and Smith 2012b), can manifest itself in many different forms, such as the integration of spoken language and posture, or images and music. As a result, multimodal approaches examine communication beyond language and focus on the full inventory of communication forms that people use (e.g. text, image, posture, gaze, etc.) and how these are interrelated. Research using these approaches relies on the premise that 'representation and communication always draw on a multiplicity of modes, all of which have the potential to contribute equally to meaning' (Jewitt 2009a: 14). The specific type of multimodality that I will focus on in this chapter is the combination of written language and visual material (i.e. text and image). Specifically, I will focus on the field of study where multimodal analysis and digital humanities intersect and examine the ways corpus-based methods can contribute to the understanding of how meaning is created through both image and text.

Research into multimodality is not a new phenomenon; the importance of multimodal means of communication has been recognised early on by a wide range of disciplines, including anthropology, psychology, sociology, art history, architecture and many more. However, multimodal research exerted a profound impact on (English) linguistics in the twentieth and twenty-first century, when, according to O'Halloran, there was an 'explicit acknowledgement that communication is inherently multimodal and that literacy is not confined to language' (2011: 123). Multimodal research was further encouraged by the advent of new technologies, both as an object of study and as a methodological tool (O'Halloran and Smith: 2012b). For instance, images have always played an important role in the way in which people interact, but it was at the beginning of the twentieth century that interest was sparked in documents and artefacts where image and text are combined. During that time, images and other graphic elements started being increasingly used in magazines and posters, a practice that has been expanded into the twenty-first century, especially with the arrival of social media (think, for example, of memes, that is, frequently reposted images or video segments on social platforms). With the advent of new technologies, new ways of

producing, distributing and interacting with images have been developed. Respectively, the means used to study images have also evolved to embrace new technologies and incorporate digital resources. However, while technology offers new possibilities, the main challenge for multimodal studies, according to O'Halloran and Smith (2012b), is to find ways to adapt the new technological resources to multimodal analysis – a challenge that echoes very strongly the main objective of digital humanities.

In order to respond to this challenge, the main question that this chapter will try to address is how we can use computational methods and tools in the investigation of multimodal data. According to Terras (2016), this is the type of question that humanities scholars have been asking themselves for some decades now, which eventually gave rise to what we recognise today as digital humanities, which can be defined as “the application of computational methods in the arts and humanities” (Terras 2016: 1637). Digital humanities is a powerful and truly interdisciplinary field; it is at the crossroads of information technology and traditional humanities research, allowing for many different types of interactions between the two (for more information on these interactions see Rosenbloom 2012).

It is important to stress at this point that considering how to use computational methods in the humanities is not something new. Scholars have used computers for a very long time to help them, for example, quantify patterns in language by using corpus-based methods. What is new is that such approaches to the study of humanities (language, arts, heritage, etc.) have, on the one hand, become much more common, while, on the other, humanities scholars are keenly interested in developing more complex digital tools, which will help them approach data from novel perspectives. Therefore, what makes digital humanities an identifiable discipline is a question of scale, both in terms of the quantity of research material studied (e.g. large-scale digitisation) and innovation (Terras 2016). This chapter echoes this focus on innovation and approaches digital humanities from the perspective of new computational tools and resources for the investigation of complex material, that is, multimodal data. With technology advancing breathtakingly quickly, we cannot be complacent about the tools we have been using for decades, but rather aim at exploring new methods that will challenge our perception of what it means to be human, not least in the way humans communicate.

Finally, an important aspect of digital humanities, which is correctly identified by Kirschenbaum (2010: 56, my emphasis), is that ‘digital humanities is more akin to a common *methodological* outlook than an investment in any one specific set of texts or even technologies’. What this methodology mainly relies on is, of course, the computer, which is a prerequisite for digital humanities research. Other definitions of digital humanities also place emphasis on the role of the computer. For instance, Burdric et al. (2012: 122, my emphasis) define digital humanities as ‘new modes of scholarship and institutional units for collaborative, transdisciplinary, and *computationally engaged* research, teaching, and publication’ (Burdric et al. 2012: 122). In the type of digital humanities research that I will focus on in this chapter, the computer is seen as a tool, which can help the researcher ‘acquire, manage, and analyse data about cultural artifacts’ (Rosenbloom 2012: online), in this case multimodal data. I will therefore explore new ways in which we can use computers to help us understand how meaning is created when image and text are combined.

Corpus-based methods fit nicely into such definitions because they engage with electronic collections (typically of texts) which can be analysed (semi-)automatically using a wide range of specifically created software tools. However, while multimodal corpora have the potential of advancing the understanding of how image and text interact, up to date, very few attempts have been made towards a corpus-based methodology which could be used for

the exploration of visual material, that is, how images contribute to meaning. The reason for this will be explored in detail in this chapter. Then, a new methodological framework for the interrogation of electronic collections of visual material will be presented which can be used in conjunction with traditional (i.e. textual) corpus approaches. Finally, this new framework will be applied to the investigation of a small multimodal corpus of popular science magazine covers in English and French.

Background

The beginning of multimodal approaches can be found in the early studies of visual material. While these did not involve the examination of text, they opened the door for multimodality and laid the foundations of what would later become recognised approaches to multimodality. Barthes was one of the first to attempt a thorough study of images with his analysis of *Paris Match* covers (1972), but it was his later work (Barthes 1977) that made a significant contribution to the study of images by distinguishing for the first time between different categories of text-image relations (i.e. anchor, illustration and relay) and applying concepts such as *denotative* (i.e. explicit or direct) and *connotative* (i.e. associated or secondary) meaning to the study of photographs. His main argument was that that contemporary images use an established lexicon to convey ideas, thus pointing towards the idea of a visual lexicon. The legacy of Barthes's work is evident by the fact that, in more recent years, these categories have been expanded into elaborate taxonomies (see for example Marsh and White 2003; Martinec and Salway 2005).

Despite the importance of Barthes's work, it was Kress and van Leeuwen (1996/2006) and O'Toole (1994/2010) who systematised the study of visual material and eventually revolutionised the study of multimodality. In their analysis of a wide range of visual material, Kress and van Leeuwen use a social semiotic approach inspired by Hallidayan grammar (1978, 1985/1994/2004), and specifically by Halliday's notion of meta-functions (ideational, interpersonal and textual). They argue that visual grammar can describe the way in which elements combine in visual statements, that is, how depicted people, places and things interact in an image to create meaning. This is similar to how textual grammars describe how words combine to form clauses. For example, they use the notion of *vector* as an equivalent of an action verb, which refers to real or virtual lines between human elements in an image. Participants in an image can be connected by a vector, in which case viewers perceive participants to be interacting (i.e. doing something to or for each other). Similarly, Kress and van Leeuwen (2006) introduce the notion of image-act which refers to how the viewer is situated in relation to the gaze of the depicted participant(s). For example, this can either be 'demand', when the participant looks directly at and demands something from the viewer, or 'offer', where such look is absent. Kress and van Leeuwen argue that their aim is to develop 'a descriptive framework that can be used as a tool for visual analysis' (2006: 14) with both analytical and critical applications.

Another scholar who uses Halliday as a starting point for the analysis of visual material is O'Toole (1994/2010). He applies central concepts and terminology of systemic-functional grammar to the study of displayed art (e.g. paintings, sculptures and architecture) and argues that 'Systemic-Functional linguistics offers a powerful and flexible model for the study of other semiotic modes besides natural language' (1994: 159). His approach differs from Kress and van Leeuwen's regarding aim. O'Toole uses artistic creation as a vantage point; his work is characterised by a close analytical study of specific visual material, focusing first and foremost on the material itself. Conversely, Kress and van Leeuwen rely much

more on the perspectives outside the material, such as historical, biographical, cultural or others, when analysing visual material. As a result, O'Toole follows a bottom-up approach, working from analyses to generalisations, while Kress and van Leeuwen advocate a top-down approach, developing first a theoretical model for the examination of visual material and analysing examples to illustrate the principles of the theoretical model. This is not to say that O'Toole disregards theoretical considerations or that Kress and van Leeuwen do not focus on analysis of material; the difference is one of starting point. This distinction between theory and data-driven approaches is still apparent in contemporary approaches to multimodal analysis.

Current contributions and research

The early approaches examined so far only focused on visual material, that is, they did not consider the interaction between image and text. In this section, I will focus on the most prominent approaches to multimodality, namely social semiotic multimodality and multimodal discourse analysis, as these are typically used for the investigation of the relationship between text and image. A brief overview of the different approaches to multimodal analysis will provide the necessary context for understanding the limitations of previous approaches and how these can be addressed using digital resources, and more specifically corpora.

Kress and van Leeuwen soon moved towards a more flexible understanding of Hallidayan grammar and expanded their work to capture multimodality (Kress 2003, 2009; Kress et al. 2001, 2004; Kress and van Leeuwen 2001; van Leeuwen 1999, 2005b), creating what is identified today as the social semiotic approach to multimodality. As with their previous work, the social semiotic approach places a strong emphasis on context, which shapes 'the resources available for meaning-making and how these are selected and designed' (Jewitt 2009b: 30), as well as on the sign-maker (i.e. the producer of the material). Any material, whether textual, visual or other, which Kress and van Leeuwen refer to as *sign*, is the result of the sign-maker's preference towards a specific resource within the given social context. For example, whether the sign-maker will decide to use a specific colour over another, or whether an image will be accompanied by text, as well as the amount of text, will be dictated by personal preferences, but also by what is considered as standard practice in any given society. In other words, the social semiotic approach to multimodality recognises that the context of production will have a strong impact on the sign. Therefore, the sign is the result of a complex interplay of the sign-maker's 'physiological, psychological, emotional, cultural, and social origins' (Kress 1997: 11), which Kress and van Leeuwen call *interest*.

A further approach to multimodality is multimodal discourse analysis (MDA), which is firmly grounded into systemic-functional grammar, following the pioneering work by O'Toole (1994/2010). The aim of this approach is to develop a comprehensive theoretical framework to describe a range of material where different semiotic resources are combined (e.g. image and text) using the idea of the Hallidayan meta-functions to understand how exactly multimodal phenomena are produced (Jewitt 2009b). Similar to research using the social semiotic approach, research using MDA (see for example Baldry 2004; O'Halloran 2004, 2005) recognises the importance of context, following the Hallidayan view that meaning is contextual. However, unlike the social semiotic approach to multimodality, MDA places very low emphasis on the sign-maker and the psychological, emotional and other origins. Also, contrary to social semiotic approaches, MDA places a higher emphasis on the *system*, defined as set choices, levels and principles in any given society.

Subsequent research examining textual and visual material focused on these two approaches (i.e. social semiotic and MDA) and expanded them further. For example, Lemke studies the interaction between visual and textual material in scientific texts (1998) and hypermedia (2002) using MDA concepts. In more recent years, multimodal research expanded rapidly as language researchers became interested in multimodality (van Leeuwen 2015). Van Leeuwen (2005a) analyses how meaning is created through complex semi-otic interactions and uses examples from photographs, adverts, magazine covers and films. While he mostly focuses on visual elements, he also refers to how these are framed by textual material (e.g. a caption in an advertisement). Martinec and Salway (2005) focus on the relationship between text and image and devise a classification for the different relations between textual and visual material. They base their classification on the Hallidayan concept of logico-semantic relations and identify, for example, the following types of *expansion* (i.e. ways of adding information): elaboration, extension and enhancement. Knox (2007) examines the visual-verbal communication on online newspaper home pages with the help of concepts from systemic-functional linguistics and critical discourse analysis. Finally, examples of some of the most recent multimodal research with text and image can be found in Ventola and Moya (2009) and Bednarek and Caple (2012).

Critical issues and topics

Limitations of digital tools

Throughout the years, with the considerable help of digital humanities, a wide range of digital tools have been made available for data collection, and it is now relatively easy to acquire a rich set of multimodal data online. Moreover, a number of data analysis tools have been developed, which are often used in the analysis of multimodal data and mostly include Computer-Assisted Qualitative Data Analysis Software (CAQDAS) (Flewitt et al. 2009), such as NVivo and HyperResearch, or annotation software, such as Praat and MacVisSta (see Rohlfing et al. 2006). Such tools help code and categorise data according to certain features (for example, NVivo allows for colour coding based on user-defined categories) and support detailed qualitative examinations of data. However, it is also important to note that these tools have not been developed with multimodal material in mind and are therefore limited in terms of what they can do. For example, MacVisSta is in principle a visualisation tool that permits the visualisation of multiple videos together with annotations.

In order to address these limitations, some attempts have been made in recent years to develop approaches that rely, at least partially, on corpus-based methods. Although not always labelled as such, these approaches belong to the realm of digital humanities as understood in this chapter, since they use computational methods in the study of humanities. However, a crucial problem when it comes to developing multimodal corpora is that corpus-based methods favour linear data, that is, data that are realised along a single dimension (Bateman 2012). This dimension is typically that of either space or time. The organisation of linear space-based data is very straightforward, which has allowed the proliferation of a large number of corpora consisting of written and spoken (i.e. transcribed) language. Time-based data are organised according to a single reference track, which is typically the speech or, more recently, the video signal, ‘since both of these are anchored in time and automatically provide orderings for their elements’ (Bateman 2008: 265). Corpora consisting of time-based data are more challenging to create due to the time-consuming process of encoding data.

Non-linear data, which are not realised along a single dimension, are typically the focus of multimodal approaches. It is clear that corpus tools developed for linear data cannot be extended to non-linear data, as they are designed with very different types of material in mind. For instance, it is not possible to use a tool, such as WordSmith, which is designed for space-based data (i.e. text) to analyse images. Therefore, new approaches and tools need to be devised, and this is where digital humanities needs to engage with innovation. The time-consuming process of developing new tools might be one of the reasons why multimodal corpora are still at an early stage of development.

Multimodal corpora

Recently, researchers have started developing such corpora which can be divided into two main categories: multimodal corpora consisting of either dynamic temporally distributed data or static spatially distributed data. The first category consists of multimodal corpora containing collections of speech and video information, which are used in research focusing on how meaning is created through speech in combination with gestures, gaze, proximity, etc. (Adolphs et al. 2011; Dybkjaer et al. 2001; Martin et al. 2005). The design of these corpora makes use of multiple annotation ‘tracks’ where different aspects of the data are recorded. For instance, a separate track is used for speech transcription, another for gesture, a third for gaze and so on (see, for example, Baldry and Thibault 2006).

The second category consists of multimodal corpora containing collections of spatially distributed static visual and textual data, which are used primarily in research focusing on how such information is organised on page or screen. This is facilitated by extending optical character recognition (OCR) technology to recognise structure and layout (Bateman 2008). Examples of the types of documents examined include patient information leaflets and other medical texts (Bouayad-Agha 2000; Corio and Lapalme 1998; Eckkrammer 2004), educational material (Walker 2003) and various printed and online documents (Baldry and Thibault 2005, 2006).

Perhaps the two most successful attempts towards a multimodal corpus tool are those by Baldry and Thibault (2008) and Bateman (2008), focusing on dynamic temporally distributed data and static spatially distributed data, respectively. Baldry and Thibault (2006) designed a toolkit based on a matrix for transcribing and analysing TV advertisements. The matrix consists of rows and columns, with each column capturing a different mode of production, for example, visual, kinesic, music, verbal, etc., and each row capturing the material (i.e. frame) for analysis. Originally, this toolkit was designed to support mostly qualitative analyses, while later Baldry and Thibault (2008) expanded it to incorporate corpus approaches by allowing searches across pre-defined parameters. In Table 6.1, frames from TV advertisements have been tagged using the matrix and then a search was conducted for the parameter [Eyes: Closed], which returned two results.

A more sophisticated tool, which does not rely on a specific theoretical framework is the GeM framework, developed by Bateman (2008; Bateman et al. 2004). Bateman follows a multi-layered approach, based on Extended Markup Language (XML), where each document is analysed according to several independent descriptions, with each of these informing a separate layer of annotation. In other words, each multimodal document, such as a page from a guidebook which consists of visual and textual elements, is divided into meaningful ‘base units’ (e.g. headings, titles, photos), and each of these units is further analysed into its components. According to Bateman, this process involves ‘defining appropriate attributes and possible values of those attributes for each type of unit introduced’ (2008: 267). This is a very promising approach which can pave the way for future multimodal analysis. However, despite

Table 6.1 Multimodal concordance

Search Feature Parameter: [Eyes: Closed]	Gaze Transitivity Frame	Verbal Text	Clause Analysis and Discourse Semantics
	GAZER^GAZE VECTOR DISENGAGED: EYES CLOSED; AVERTED DOWNWARDS: DISENGAGE INTERLOCUTOR	<i>None</i>	No clause analysis; Discourse Move: Signals inability to provide desired response
	GAZER^GAZE VECTOR DISENGAGED: EYES CLOSED; AVERTED DOWNWARDS: DISENGAGE INTERLOCUTOR	<i>You sat opposite the getaway car!</i>	Exp: Actor^Process: Material: Action^Circumstance: Location; Int: declarative; second person Subject; Tex: Theme/Given: second person focus; Rheme/New: end focus on addressee's close proximity to getaway car. Discourse Move: Signal Incredulity, surprise

Exp: = Experiential; Int: = Interpersonal; Tex: = Textual (Metafunctions)

^ = followed by (see Halliday/Matthiessen 2004 [1985])

the considerable potential of the GeM framework, Bateman does not explain in detail how visual elements can be further analysed once they have been identified. Research with visual material (see later sections) has demonstrated that the visual channel is a complex semiotic mode, and this complexity should not be ignored if we want to fully capture multimodality.

Therefore, with the exception of the multimodal concordance tool designed by Baldry and Thibault (2008), and Bateman's (2008) GeM framework, most other tools, while allowing us to approach data from novel perspectives, do not seem to support thorough quantitative analyses. What seems to be missing is a principled and structured methodology that can allow for quantitative analyses of multimodal material. This methodology does not need to be necessarily restricted to a specific linguistic or other theoretical framework, as is the case with Baldry's and Thibault's toolkit, which relies heavily on systemic-functional grammar. It also needs to provide, unlike the GEM framework, sufficient room for the detailed interrogation of visual material. Thus, while innovation is evident in previous research, substantial developments still need to be made in this area.

Challenges for multimodal corpora

Such developments are, however, halted by the fact that different modes have their own analytical challenges and limitations, both individually and in combination. Some material (e.g. audio-visual or hypertext) presents difficulties when it comes to annotation, as it is difficult to capture their dynamism and multiple dimensions (O'Halloran and Smith 2012a). Regarding the visual mode, one of the challenges typically identified is how to 'tag' images to allow for meaningful searches. This has naturally resulted in multimodal research being small-scale, focusing on the detailed examination of a handful of instances. According to Jewitt (2009a) 'the scale of multimodal research can restrict the potential of multimodality to comment beyond the specific to the general'. She argues that these limitations might be overcome by the development of multimodal corpora, as well as by using quantitative methods in multimodal analysis, which can, in principle, capture large quantities of data (for example thousands of images).

However, the problem of how to capture the dynamic and complex nature of some media is still present even with corpus-based approaches. Therefore, the challenge for digital humanities is to find ways in which existing principles of corpus-based methods can be informed by and enriched with elements that will allow for the (semi-)automated computational analysis of multimodal texts. In the case of material combining image and text, which is the focus of this chapter, what is needed is an appropriate mark-up scheme for the investigation of visual material, which can inform corpus-based approaches, facilitate the annotation process and guide analysis. The mark-up scheme needs to rely on clearly defined parameters, which at the same time remain flexible, to accommodate the aims of different research projects. Developing a mark-up scheme is an essential step in devising a methodology that can allow for potentially large-scale quantitative analyses of visual corpora. The idea of attributes and values, introduced by Bateman (2008), can prove particularly useful when designing this mark-up scheme.

Main research methods

Visual content analysis

As explained earlier, if corpora of visual material are to be developed, any mark-up scheme we create needs to be able to capture the complexity of visual material, and at the same time,

be compatible with corpus-based methods. I propose here content analysis as a methodological framework that can inform the scheme for visual analysis. According to Bell (2011), content analysis is ‘an empirical (observational) and objective procedure for quantifying recorded “audio-visual” (including verbal) representation using reliable, explicitly defined categories (“values” on independent “variables”’. While content analysis is not restricted to visual material, Bell proposes a framework which he names visual content analysis and applies some of its principles to the investigation of a small collection of women’s magazine covers.

To allow for a systematic and meaningful analysis, content analysis begins with a specific hypothesis or question about well-defined variables. Variables refer to specific dimensions, such as size, colour, position, participants, etc., and are logically and conceptually independent of each other. Each variable consists of observable categories, that is, values, which are mutually exclusive and exhaustive (in some cases that means that a *miscellaneous* category needs to be included). For example, a valid hypothesis might be that women are depicted more often than men in women’s magazine covers. Therefore, the depicted participant can be identified as a relevant variable, with male and female as values.

The more complex the material that needs to be analysed, the greater the importance of developing precise definitions of variables and values. While physical features (such as size and colour) might generally be easy to define, other features might be more ambiguous (such as age and race) and therefore need to be avoided or at least be approached with caution. According to Bell (2011), this is because each of these features consists of a number of more specific features. Also, some variables, such as participants or colour, might take more than one value, for example, if an image depicts both a man and a woman or uses different colours. Therefore, researchers need to focus on the variable and values that are most relevant to the research project in question, avoiding composite variables that might create difficulties in coding.

It is also important to note that content analysis is used for comparative hypotheses or questions, rather than to calculate absolute frequencies. The aim of visual content analysis is to test the ways in which the media represent people, objects and events by allowing ‘quantification of samples of observable content classified into distinct categories’ (Bell 2011: 14). It is an empirical, systematic and objective procedure which relies on the identification of observable dimensions of the visual material in question and judgements about how frequently certain relevant features appear compared to others.

Content analysis has been criticised (see Hall 1980; Winston 1990) for not being sufficiently informed by relevant theories since it focuses mostly on common-sense categories that can be easily identified, such as gender or size. As a result, such analyses can also be decontextualised, ignoring some aspects of the semiotic codes of visual material, such as the historical, cultural or social context in which an image is created and distributed, which have been very important in both social semiotic approaches to multimodality and MDA. For example, the assumptions that the audience has about the publication in which an image appears will naturally have an impact on how they interpret it, creating each time a different effect, even though the image might remain more or less the same. The same central image in a tabloid or a spreadsheet newspaper is bound to be perceived and interpreted differently, even by the same people in the same social and cultural context. Needless to say, that even more variation is introduced once we consider people with different social and cultural backgrounds. A way of addressing these limitations is by allowing relevant theories and evidence from the sociocultural contexts to inform the analysis. While these limitations are valid, they do not compromise the significant descriptive and methodological potential of

visual content analysis. Therefore, the aim here is not to engage in a discussion or address them, but rather to offer an example of how visual content analysis can inform the creation of multimodal corpora.

From visual content analysis to visual corpora

The principles of visual content analysis can easily be expanded to fit the aims and principles of corpus-based methods and inform the mark-up scheme for the creation of visual corpora. Each image to be included in the corpus can be defined in terms of relevant variables and values, based on the research questions of the project, which can then be converted into tags. A corpus can then be searched based on one or more tags and categories can be quantified. An example of a list of possible variables (i.e. participant, posture, setting and size) and their associated values can be found in Table 6.2. Based on this, it is possible to tag an image in a magazine depicting a man sitting on rock with the following tags: human, sitting, rural, full page.

The rationale behind this mark-up scheme is very similar to part-of-speech tagging for linguistic data. However, while parts of speech are generally accepted as universal categories, the categories for visual tagging need to be carefully identified and defined based on the nature of each individual project. Although Bell (2011) argues that values need to be mutually exclusive, I believe that this cannot, and should not, always be the case. A typical example would be a children’s book cover, such as the *Red Riding Hood*, depicting both a human (i.e. a little girl) and an animal (i.e. a wolf). This might become even more complicated if other human participants are included (i.e. the grandmother and the hunter). In such cases, which are not infrequent, I argue that an image can be tagged with more than one value from each variable, or alternatively, that each participant is tagged separately, therefore breaking down the image into its main constituents. An alternative would be to focus on the principal participant if it can be easily identified and is in accordance with the aim of the research project.

While this might create the impression that visual content analysis is rather complex, in reality, the research question(s) will guide most of the analysis. In particular, since visual content analysis is used for comparative purposes, it will typically involve only a few variables related to the aim of the project. The aim is to tag images based on relevant variables and not attempt an exhaustive description. For instance, a hypothetical research question might be whether humans or animals appear more frequently on children’s book covers. In this case, only the participant variable is required. An additional research question can be added, namely whether these participants appear in a domestic or rural setting, which can be captured by the setting variable. As more questions are added to the project, more variables, and their associated values, become relevant. Analytically, such a scheme presupposes the

Table 6.2 Hypothetical values on four variables

Variables				
Values	Participant	Posture	Setting	Size
	Human	Sitting	Urban	Full page
	Animal	Standing	Rural	Half full page
	No participant			

creation of a relevant tool that can be used for the tagging and investigation of corpora consisting of visual material. Although simple desktop applications, such as spreadsheets, can be used to some extent for this purpose, more sophisticated tools are needed, such as the one that will be presented in the next section.

Sample analysis

Data and methods

In the remainder of this chapter, I will present a brief case study investigating how science, which is understood here as the study of the physical world, is depicted in the magazine covers of an American and a French popular science publication. According to Kress and van Leeuwen (1996/2006) images are not universal and there is likely to be regional and social variation in visual design. As far as European countries are concerned, they argue that '[s]o long as the European nations and regions still retain different ways of life and a different ethos, they will use the 'grammar of visual design' distinctly' (Kress and Van Leeuwen 2006: 5). This proposition will be tested by comparing the magazine covers of *Scientific American* and *La Recherche*, two popular science magazines. The focus is on the central image and the main headline of the magazine covers accompanying that image.

The specific research questions of the project are the following: (1) whether differences can be observed between *Scientific American* and *La Recherche* in the proportion of animate and inanimate participants depicted in the central image of the magazine cover, (2) whether *Scientific American* and *La Recherche* demonstrate different preferences with regard to the most frequent animate and inanimate participants, and (3) whether there are any differences between *Scientific American* and *La Recherche* in the distribution of nouns and verbs (both overall proportions and most frequent items) in the main headline, which accompanies the central image of the magazine cover. Ultimately, by combining answers to these questions, the analysis will reveal whether, and to what extent, the two publications represent science differently through both image and text and how exactly science is presented and communicated to the public.

While naturally the two magazines do not focus on the same stories (although some overlap is observed), this does not present a limitation, since the aim here is not to compare how, for instance, the English and French refer to space exploration, but rather, and perhaps more importantly, whether one of the two publications shows a stronger preference towards space (compared to animals, for example). Therefore, comparisons are achieved by relying on the same variables and values, even though their distribution is expected to be different, with the two publications showing different preferences. Given the general lack of corpus-based methods for the analysis of visual material, as well as the methodological outlook of digital humanities, more emphasis is placed here on discussing the methodology and tools developed for the examination of the visual rather than the linguistic components.

Data consist of visual and textual elements taken from the covers of *Scientific American* and *La Recherche* from the period 2005–2014. These popular science magazines were selected as they are well-known and circulate widely in their respective countries. Overall, 239 images were collected (120 for *Scientific American* and 112 for *La Recherche* – *La Recherche* is published monthly, apart from the period July–August, but this changed in 2013, when separate issues circulated for each of the two months) and 2,511 words (1,545 for *Scientific American* and 966 for *La Recherche*). Words here refer to the main title accompanying the image under investigation. It is already obvious that the French magazine

prefers shorter titles compared to the English one (on average, English headlines consist of 13 words, while French of 9). This might suggest that the French publication relies more on the visual than the textual mode to convey a message compared to the English one.

Corpus size and scope are important parameters in any corpus-based project, and while previous research in corpus linguistics tends to favour large (i.e. millions of words) corpora, there is an ever-increasing amount of work that focuses on small-scale, bespoke corpora which are constructed to answer specific research questions, suggesting that decisions regarding corpus size are ultimately project-specific. The corpus used for this case study is considered small, especially as far as the textual component is concerned. However, the main aim here is to showcase how a multimodal corpus-based methodology, which is based on visual content analysis, can be applied to empirical data, as well as the type of questions that can be addressed using it. Therefore, the corpus size is considered adequate for this aim to be achieved. Methodologically, a multimodal corpus is analysed in two stages. The first stage of analysis involves the examination of the visual corpus components (i.e. central images), while the second stage consists of the textual components (i.e. main headlines).

During the first stage of analysis, the principles of visual content analysis are applied to the examination of a corpus of images from the magazine covers. The focus is on the depicted participant of the magazine cover, which can take many forms (e.g. a human, a robot, an animal, an object, a landscape, etc.). Analysis is divided into two parts, each of which addresses a separate research question. First, the focus is on the proportions of animate and inanimate participants. The category of animate participants captures anything that can be associated with living organisms and includes covers depicting human and human-like beings, body parts (e.g. eye or hand), animals, body organs (e.g. brain) and the microcosm (e.g. cells, viruses and bacteria). Human-like beings, such as robots or androids, are included as these are often used symbolically to refer to human properties. The inanimate category includes objects, shapes, landscapes and anything else that cannot be categorised as animate. Therefore, the first part of the analysis of the visual component is based on only one parameter (i.e. the participant), and it is possible to choose from two values (i.e. *animate* or *inanimate*).

Second, the focus is on the proportions of specific frequent animate and inanimate participants. Pilot studies analysing the visual components indicate that the following participants typically appear in at least one of the publications: *animals*, *microcosm* (e.g. anything that can only be perceived using a microscope, such as cells, neurons, bacteria and viruses) and *human brain* for animate participants and *space* (e.g. stars and planets other than Earth) and *Earth* (either landscapes on Earth or images of the planet from space) for inanimate participants. Therefore, the second part of the analysis of the visual components is once again based on one parameter (i.e. the participant), but there are now altogether five values to allow for a more detailed analysis (Figure 6.1). In cases where two or more participants are depicted that belong to different values, the focus is on the most dominant participant.

To facilitate tagging and analysis, a purpose-built software, Visual Tagger, which runs on a dynamic web page, whose content can be controlled by the user, and which has been developed using Hypertext Preprocessor (PHP), Hypertext Markup Language (HTML), and JavaScript, is used. The software saves data on a MySQL database, an open-source relational database management system. The software allows for unlimited tags to be added to each image, which can then be used as the basis for search queries. It also allows images to be saved in separate folders, an important feature when comparisons between different publications need to be made. Finally, Visual Tagger allows for images to be tagged based on the date they were produced, adding another useful parameter for comparisons. Each image

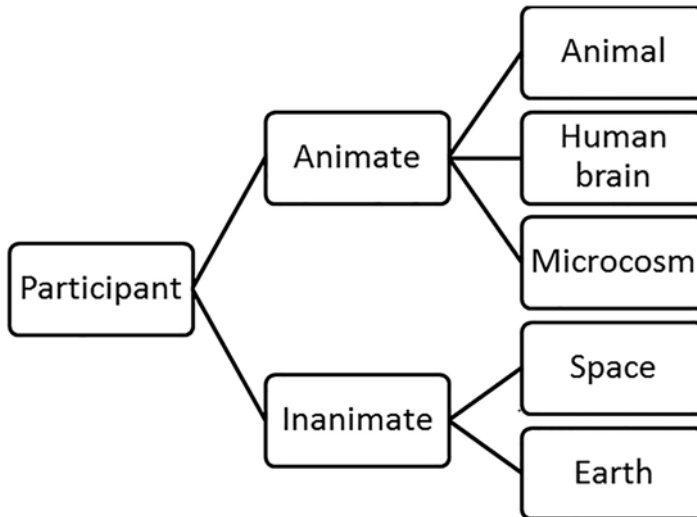


Figure 6.1 Parameter and values

is opened using Visual Tagger and a tag is added based on the values mentioned earlier. The images are saved in a separate folder, one for each popular science magazine (Image 6.1). The use of this tool, together with the mark-up scheme, is a good example of how the work that digital humanists do is at the start of the innovation curve (Terras 2016).

Once all images have been tagged, the search facility of Visual Tagger is used to search for tags within each folder (Image 6.2).

Finally, all instances of a tag are calculated for each magazine. The software is at an early stage of development, with more features planned to be developed in the future. Therefore, at the moment, the software is not in the public domain and cannot be freely accessed.

For the second stage of analysis, which involves the analysis of the textual components, Sketch Engine is used, which is a widely used online corpus management and analysis tool. Two sub-corpora are created: one consisting of the main headlines of *Scientific American* and the other of the main headlines of *La Recherche*. The sub-corpora are then tagged for parts of speech, using the in-built tagger of Sketch Engine. Finally, separate searches are conducted for each of the sub-corpora based on the verb and noun tags. These categories were selected for two main reasons. On the one hand, they appear more frequently in headlines compared to, for example, adjectives and adverbs (Molek-Kozakowska 2014). On the other, the analysis of nouns can also be associated with the type of participants identified during the examination of the visual components. Wildcards are used to maximise the number of instances captured, for example [tag= "N.*"] captures singular nouns (NN), plural nouns (NNS), possessive nouns (NNSZ), etc. The instances are recorded and proportions of nouns and verbs (e.g. percentages) are calculated. This second stage of analysis therefore aims to address the final research question, that is, whether there are any differences between *Scientific American* and *La Recherche* in the distribution of nouns and verbs (both overall proportions and most frequent items) in the main headline, which accompanies the central image of the magazine cover. While this methodology is fairly simple, it serves to demonstrate how digital humanities, through the use of corpus methods, can facilitate multimodal analyses, especially where the focus is on how meaning can be created through both image and text.



Image 6.1 Screenshot of Visual Tagger tagging interface



Image 6.2 Screenshot of Visual Tagger search interface

Results from the visual components

The first stage of analysis involved the examination of the visual components (i.e. magazine covers) with regard to the central participant. First, each image was tagged using one of the following values: *animate* and *inanimate*. Results reveal similar patterns for *Scientific American* and *La Recherche* (Table 6.3).

Overall, animate and inanimate participants are equally distributed in the covers of each publication. However, *Scientific American* shows a slightly stronger preference for inanimate compared to animate participants, while *La Recherche* follows the reverse pattern. The difference in the proportions is too small, and, at least on this broad level of analysis, the two cultures seem to employ the ‘grammar of visual design’ similarly.

Second, the images within each category (i.e. animate and inanimate) were tagged for the most frequent types of participants using the following values: *animals*, *human brain* and *microcosm* for the animate category, and *space* and *Earth* for the inanimate category. Based on this analysis, some more noticeable differences can be observed (Table 6.4).

None of the covers of *Scientific American* depicting an animate participant included animals, while four (6.8 percent) instances of animals were captured in *La Recherche*. While the proportion is very small to allow for meaningful patterns to be observed, it can be noted that two out of these four covers depicted dinosaurs. A different distribution was also observed regarding the human brain, which was found to be a popular theme for both publications. This theme was recorded 14 (25.0 percent) times in *Scientific American*, and 8 times (13.6 percent) in *La Recherche*. Finally, differences were observed in the microcosm category, that is, anything that can only be perceived with a microscope. *La Recherche* shows a stronger preference for this category, with 12 covers using this theme (20.4 percent), while only 6 (10.7 percent) were found in *Scientific American*. Based on these differences, it appears that *Scientific American* shows a stronger preference for themes more closely associated with human beings compared to *La Recherche*.

Table 6.3 Distribution of animate and inanimate participants in *Scientific American* and *La Recherche*

	<i>Scientific American</i>		<i>La Recherche</i>	
	<i>Raw frequency</i>	<i>Normalised frequency</i>	<i>Raw frequency</i>	<i>Normalised frequency</i>
Animate	56/120	46.7%	59/112	52.7%
Inanimate	64/120	53.3%	53/112	47.3%

Table 6.4 Distribution of specific animate and inanimate participants in *Scientific American* and *La Recherche*

	<i>Scientific American</i>		<i>La Recherche</i>	
	<i>Raw frequency</i>	<i>Normalised frequency</i>	<i>Raw frequency</i>	<i>Normalised frequency</i>
Animals	0	0.0%	4/59	6.8%
Human brain	14/56	25.0%	8/59	13.6%
Microcosm	6/56	10.7%	12/59	20.4%
Space	19/64	29.7%	9/53	17.0%
Earth	12/64	18.7%	4/53	7.6%

Regarding inanimate participants, *Scientific American* shows a very strong preference towards the space theme, with 19 covers (29.7 percent) depicting stars and planets other than Earth. *La Recherche* uses this theme less frequently, namely only nine covers (17.0 percent) were found depicting space. A similar pattern is observed regarding the Earth theme. *Scientific American* shows a stronger preference, with 12 covers depicting images of Earth (18.7 percent), while *La Recherche* uses this theme only 4 times (7.6 percent). If we consider that the Space–Earth distinction might also symbolically represent inclusion and exclusion of humankind (i.e. Earth includes humankind and Space excludes humankind), there is evidence to suggest that while both publications focus more on exclusion, *La Recherche* generally tends to avoid more strongly depictions of humankind. This is in line with the findings from the analysis of animate participants, where *Scientific American* was found to prefer human-related themes more than *La Recherche*. This difference in preferences for visual depictions of science could also be explained by the individualism–collectivism dimension proposed by Hofstede (2001). However, more data are needed if this hypothesis is to be tested.

Results from the textual components

The second stage of analysis involved the examination of the textual components in terms of the proportions of nouns and verbs used in the main headlines of *Scientific American* and *La Recherche* accompanying the visual material studied earlier (Table 6.5). Nouns include singular and plural nouns, possessive nouns and proper names. Verbs include all verb forms, as well as participles and gerunds, but not modal verbs, as they are typically part of a verb phrase. Since many verb phrases in both English and French can consist of more than one verb (e.g. *have been eating*), results are manually refined, which is facilitated by the small size of the sub-corpora. It should be noted that, based on data obtained from the analysis of enTenTen and frTenTen (an English and a French corpus of texts collected from the Internet, which belong to the TenTen corpus family and consist of over 10 billion words each) both languages employ nouns and verbs with a similar frequency: nouns represent 21.13 percent of enTenTen and 19.44 percent of the frTenTen, while verbs are only slightly more frequent in the English corpus (17.15 percent) than the French one (13.71 percent). This comparison suggests that the typical frequencies of nouns and verbs in these two languages are comparable and that any differences that might be observed in the examination of the textual corpus components need to be attributed to the specific genre and text type, that is, popular science headlines.

It was found that *Scientific American* and *La Recherche* employ nouns with approximately the same frequency – 3657 and 3675 per 10,000 words respectively (0.5 percent

Table 6.5 Distribution of verbs and nouns in *Scientific American* and *La Recherche*

	<i>Scientific American</i>		<i>La Recherche</i>	
	<i>Raw frequency</i>	<i>Normalised frequency (per 10,000 words)</i>	<i>Raw frequency</i>	<i>Normalised frequency (per 10,000 words)</i>
Nouns	565/1,545	3,657	355/966	3,675
Verbs	196/1,545	1,269	77/966	797

difference). Thus, in both languages, nouns correspond to approximately one-third of all words in the main headlines. This suggests a general preference towards nouns in both languages, which can be justified by the nature of the publication (popular science) and the text type (headline). Indeed, the proportions of nouns here are much higher than those found in the TenTen corpora. Previous research (Jenkins 1987; Bucari 2004; Marcoci 2014; Bednarek and Caple 2012) has shown that headlines show a very strong preference towards nouns; a phenomenon that Jenkins (1987, p. 349) calls ‘stacked nouns’. The limited space available to headlines, together with the fact that they need to attract readers’ attention, is the reason behind this. The most frequent nouns in English are *future*, *life* and *universe*, while the most frequent nouns in the French headlines are *science*, *cerveau* (brain) and *vie* (life) (see Examples 1 and 2).

(1) Comment notre cerveau se répare, se remodèle, se régénère
[*La Recherche* July–August 2007]

How our brain repairs, remodels, regenerates itself
[near-literal translation]

(2) Hidden worlds of dark matter: An entire universe may be inter-woven silently with our own
[*Scientific American* November 2010]

Therefore, while both publications widely employ nouns, they discuss slightly different topics, with French focusing more on human-related concepts, such as the brain. This is further supported by the examination of proper nouns, such as *Einstein*, *Peter Higgs* and *Steve Jobs*. *Scientific American* uses 4 proper nouns referring to people, while *La Recherche* uses 14. This corresponds to 26 per 10,000 words in the English sub-corpus and 145 per 10,000 in French sub-corpus (139 percent difference). Therefore, as far as headlines are concerned, there is an indication that the French texts might show a preference towards concepts associated with humans compared to the English ones. This contrasts with the findings from the previous stage of analysis. However, the preference of *Scientific American* for headlines referring to *universe* is in line with what was found during the examination of the visual components.

Regarding verbs, it was found that both publications employ very few verbs in headlines; fewer than what is found in the TenTen corpora. This is to be expected, since ellipsis is one of the most common features of headlines (Jenkins 1987; Bell 1991; Reah 1998) and a study conducted by Molek-Kozakowska (2014) found that the headlines of the popular science magazine *New Scientist* employed a very low number of verbs. Comparing the two publications, *Scientific American* employs more verbs than *La Recherche* – 1269 and 797 per 10,000 words respectively (45.7 percent difference). This might be related to the tendency of *Scientific American* to use longer headlines and incorporate long subtitles, which increases the chances of verbs appearing. However, this is not necessarily the case, as Example 1 from *La Recherche* demonstrates. Compared to nouns, verbs are used considerably less frequently in both publications, but the difference is more marked in the case of *La Recherche*. The most frequent verbs in English apart from *be* and *have* are *change*, *make* and *get*, while the most frequent verbs in French, apart from *être* (be) and *avoir* (have) are *changer* (change), *révéler* (reveal) and *exister* (exist). There are some similarities in these choices, and *change* is used frequently in both languages. However, *La Recherche* also demonstrates a clear tendency for verbs related to discovering (Example 3), which is not observed in *Scientific American*.

(3) Comment nous devenons intelligents: Ce que révèle l'imagerie cérébrale
[*La Recherche* November 2011]

How we become intelligent: What brain scans reveal
[near-literal translation]

This is further supported by the fact that the noun *decouvertes* (discoveries) appears five times in the French sub-corpus. The focus of the French publication on discoveries might also be related to its preference for the microcosm theme, identified during the previous stage of analysis.

Discussion

The analysis suggests that there are both similarities and differences between *Scientific American* and *La Recherche* in the way science is presented through the textual and visual mode. Regarding the visual mode, significant similarities are observed in the proportions of animate and inanimate participants. However, differences are observed between the two publications with regard to the most frequent animate and inanimate participants. These involve a stronger preference towards images of the human brain, outer space and Earth by *Scientific American* compared to *La Recherche* and a stronger preference for images of animals and the microcosm, as well as a weaker emphasis on images of Earth by *La Recherche* compared to *Scientific American*. Therefore, there is an initial indication that *Scientific American* places more emphasis on humans and the world they inhabit, but at the same time, it also considers the larger framework in which humans exist (i.e. outer space). This creates a unique balance through visual means between the different themes. Such a balance is not observed in *La Recherche* where emphasis is clearly placed on non-human-oriented concepts. It is also evident that *La Recherche* focuses more on the micro-level, while *Scientific American* on the macro-level, as the results from the microcosm tag suggest.

Regarding the textual mode, both publications show a stronger preference towards nouns instead of verbs (ellipsis), while the proportions of nouns are also similar. In terms of differences, *La Recherche* was found to clearly favour nouns associated with humans and verbs associated with discovery. A similar tendency was not observed in *Scientific American*. The balance achieved through visual means in *Scientific American* is strengthened through textual means, where less emphasis is placed on human-related concepts compared to the visual channel. A similar balance is achieved in *La Recherche* where the weak emphasis on human-related concepts in the visual mode is compensated in the textual mode. Therefore, this should not be perceived as a contradiction, but rather as a way of combining semiotic modes to reinforce different aspects of the same message. The depiction of science in these magazine covers does not rely on a single mode (textual or visual), but on both, and each publication exploits these two modes differently to achieve a certain balance and communicate the message to the reader. Due to the complexity of this balance, it is difficult to arrive at a valid conclusion about any differences between the two publications in terms of representations of science. Both publications are concerned with similar themes and topics, but these are communicated using different modes. Therefore, although these two publications might use the 'grammar of visual design' differently, the message they convey is similar when the visual and textual modes are combined.

Future directions

A relatively brief analysis of material from popular science publications has shown that digital humanities, understood here as the combination of information technology and traditional humanities research, can offer a more complete understanding of how messages are created and communicated both textually and visually through the development and interrogation of multimodal corpora in this particular case study. This is achieved by focusing on pre-defined relevant semiotic modes (i.e. visual and textual). Had I focused my analysis on only one of the semiotic modes, I would have reached very different conclusions and would have perhaps been able to identify significant differences between the two publications. However, it should be clear by now that this would have offered a limited, and potentially skewed, understanding of the phenomenon.

Although the corpora examined in this case study are rather small, corpus-based methods allow for much larger corpora to be built and analysed. Relying on large amounts of data and examining these quantitatively can significantly increase the validity of the conclusions. This is common practice for textual data, but still a novel approach for visual data. Therefore, digital humanities, in the form of corpus-based methods, can, and should, be expanded to develop methodologies and frameworks which can reveal the meaning-making dynamics of all semiotic resources available. In the future, we are likely to see more mark-up schemes being developed which are sophisticated, yet easily accessible to humanities scholars, and which can inform corpus analysis. This will be an indication that we are moving from the stage of innovation, on which this chapter has focused, to adoption.

Subsequently, we are likely to see more specifically designed tools being created and used more widely, similar to those already used for textual corpora (e.g. WordSmith Tools, Sketch Engine, Antconc). In the same way as monolingual corpora are currently offering different perspectives into law, history, media, social science and other disciplines, multimodal corpora are likely to change our perception of the text–image interaction in the future. Although research with such corpora is still at an early stage, due to the inherent difficulties associated with the complex nature of multimodality, this is an avenue of investigation that must be further encouraged. Ultimately, what I hope we will see in the future is more research with multimodal corpora, whether these combine visual and textual material or posture, gaze, gestures, intonation, music, layout, etc., which will revolutionise our understanding of how different modes interact to create meaning.

Further reading

- 1 Bell, P. (2011). Content analysis of visual images. In T. van Leeuwen and C. Jewitt (eds.), *Handbook of visual analysis*. London: Sage Publications, pp. 10–34.
This chapter introduces the main principles of visual content analysis, which is based on the idea of variables and values, and applies it to the examination of women's magazine covers.
- 2 Baldry, A. and Thibault, P.J. (2006). *Multimodal transcription and text analysis*. London: Equinox. This book investigates in-depth issues pertaining to the analysis of multimodal texts, such as how to transcribe and analyse them. It investigates a range of material, such as printed texts, websites, and films, and offers solutions to the current problems of multimodal text analysis and transcription.
- 3 Bateman, J. (2008). *Multimodality and genre: A foundation for the systemic analysis of multimodal documents*. Hampshire: Palgrave Macmillan.
This book presents the first systematic attempt towards a corpus-based approach to the description and analysis of multimodality. It relies on principles of multimodal discourse analysis and offers a detailed presentation of the GeM framework for the analysis of multimodal texts.

- 4 Knight, D. (2011). The future of multimodal corpora. *Revista Brasileira de Linguística Aplicada* 11(2): 391–415.

This article offers a critical overview of existing approaches to multimodal corpus analysis, that is, multimodal corpora that have been constructed in the past, and discusses the technological and methodological innovations that can help advance multimodal corpora in the future.

References

- Adolphs, S., Knight, D. and Carter, R. (2011). Capturing context for heterogeneous corpus analysis: Some first steps. *International Journal of Corpus Linguistics* 16(3): 305–324.
- Baldry, A. (2004). Phase and transition, type and instance: Patterns in media texts as seen through a multimodal concordancer. In K.L. O'Halloran (ed.), *Multimodal discourse analysis: Systemic functional perspectives*. London: Continuum, pp. 83–108.
- Baldry, A. and Thibault, P.J. (2005). Multimodal corpus linguistics. In G. Thompson and S. Hunston (eds.), *System and corpus: Exploring connections*. London and New York: Equinox, pp. 164–183.
- Baldry, A. and Thibault, P.J. (2006). *Multimodal transcription and text analysis*. London: Equinox.
- Baldry, A. and Thibault, P.J. (2008). Applications of multimodal concordances. *Hermes – Journal of Language and Communication Studies* 41: 11–41.
- Barthes, R. (1972). *Mythologies* (A. Lavers, trans.). London: Jonathan Cape.
- Barthes, R. (1977). *Image, Music, Text* (S. Heath, trans.). London: Fontana Press.
- Bateman, J. (2008). *Multimodality and genre: A foundation for the systemic analysis of multimodal documents*. Hampshire: Palgrave Macmillan.
- Bateman, J. (2012). Multimodal corpus-based approaches. In C. Chappell (ed.), *The encyclopedia of applied linguistics*. Hoboken, MA and Oxford: Blackwell Publishing Ltd.
- Bateman, J., Delin, J. and Henschel, R. (2004). Multimodality and empiricism: Preparing for a corpus-based approach to the study of multimodal meaning-making. In C. Ventola and M. Kaltenbacher (eds.), *Perspectives on multimodality*. Amsterdam and Philadelphia: John Benjamins, pp. 65–87.
- Bednarek, M. and Caple, H. (2012). *News discourse*. London: Continuum.
- Bell, A. (1991). *The language of news media*. Oxford: Blackwell.
- Bell, P. (2011). Content analysis of visual images. In T. Van Leeuwen and C. Jewitt (eds.), *Handbook of visual analysis*. London: Sage Publications, pp. 10–34.
- Bouayad-Agha, N. (2000). Layout annotation in a corpus of patient information leaflets. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhaouer (eds.), *Proceedings of the second international conference on language resources and evaluation*. Athens, Greece: European Language Resources Association (ELRA).
- Bucaria, C. (2004). Lexical and syntactic ambiguity as a source of humour: The case of newspaper headlines. *Humor* 17(3): 279–309.
- Burdrick, A., Drucker, J., Lunenfeld, P., Presner, T. and Schnapp, J. (2012). *Digital humanities*. Cambridge, MA: MIT Press.
- Corio, M. and Lapalme, G. (1998). *Integrated generation of graphics and text: A corpus study*. In *Proceedings for the COLING-ACL Workshop on Content Visualisation and Intermedia Representations* (pp. 63–68). Montreal, Canada: Association for Computational Linguistics.
- Dybkjaer, L., Berma, S., Kipp, M., Olsen, M.W., Pirreli, V., Reithinger, N. and Soria, C. (2001). *Survey of existing tools, standards and user needs for annotation of natural interaction and multimodal data, international standards for language engineering (ISLE): Natural interactivity and multimodality working group deliverable*. Odense, Denmark: NISLab, Odense University.
- Eckkrammer, E.M. (2004). Drawing on theories of inter-semiotic layering to analyse multimodality in medical self-conselling texts and hypertexts. In E. Ventola, C. Charles and M. Kaltenbacher (eds.), *Perspectives on multimodality*. Amsterdam and Philadelphia: John Benjamins, pp. 211–226.
- Flewitt, R., Hampel, R., Hauck, M. and Lancaster, L. (2009). What are multimodal data and transcription? In C. Jewitt (ed.), *The Routledge handbook of multimodal analysis*. London and New York: Routledge, pp. 40–53.

- Hall, S. (1980). Encoding/Decoding. In S. Hall, D. Hobson, A. Lowe and P. Willis (eds.), *Culture, media, language*. London and New York: Routledge, pp. 117–127.
- Halliday, M.A.K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. London: Edward Arnold.
- Halliday, M.A.K. (1985/1994/2004). *An introduction to functional grammar*. London: Edward Arnold.
- Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations*. Thousand Oaks, CA: Sage Publications.
- Jenkins, H. (1987). Train sex man fined: Headlines and cataphoric ellipsis. In M.A.K. Halliday, J. Gibbons and H. Nicholas (eds.), *Learning, keeping, and using language*. Amsterdam: John Benjamins, pp. 349–362.
- Jewitt, C. (2009a). An introduction to multimodality. In C. Jewitt (ed.), *The Routledge handbook of multimodal analysis*. London and New York: Routledge, pp. 14–27.
- Jewitt, C. (2009b). Different approaches to multimodality. In C. Jewitt (ed.), *The Routledge handbook of multimodal analysis*. London and New York: Routledge, pp. 28–39.
- Kirschenbaum, M. (2010). What is digital humanities and what's it doing in English departments? *ADE Bulletin* 150: 55–61.
- Knight, D. (2011). The future of multimodal corpora. *Revista Brasileira de Linguística Aplicada* 11(2): 391–415.
- Knox, J. (2007). Visual-Verbal communication in online newspaper home pages. *Visual Communication* 6(1): 19–53.
- Kress, G. (1997). *Before writing: Rethinking paths to literacy*. London and New York: Routledge.
- Kress, G. (2003). *Literacy in the new media age*. London and New York: Routledge.
- Kress, G. (2009). *Multimodality: A social semiotic approach to communication*. London and New York: Routledge.
- Kress, G., Jewitt, C., Bourne, J., Franks, A., Hardcastle, J., Jones, K. and Reid, E. (2004). *English in urban classrooms: Multimodal perspectives on teaching and learning*. London and New York: Routledge.
- Kress, G., Jewitt, C., Ogborn, J. and Tsatsarelis, C. (2001). *Multimodal teaching and learning: The rhetorics of the science classroom*. London: Continuum.
- Kress, G. and Van Leeuwen, T. (1996/2006). *The grammar of visual design*. London and New York: Routledge.
- Kress, G. and Van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. London: Edward Arnold.
- Lemke, J.L. (1998). Multiplying meaning: Visual and verbal semiotics in scientific text. In J.R. Martin and R. Veel (eds.), *Reading science: Critical and functional perspectives on discourses of science*. London and New York: Routledge, pp. 87–113.
- Lemke, J.L. (2002). Travels in hypermodality. *Visual Communication* 1(3): 299–325.
- Marcoci, S. (2014). Some typical linguistic features of English newspaper headlines. *Linguistic and Philosophical Investigations* 13: 708–714.
- Marsh, E.E. and White, M. (2003). A taxonomy of relationships between images and text. *Journal of Documentation* 59(6): 647–672.
- Martin, J.-C., Kuhnlein, P., Paggio, P., Stiefelhagen, R. and Pianese, F. (eds.) (2005). *Proceedings of workshop on multimodal Corpora: From multimodal behaviour theories, to usable models*. Genova, Italy: LREC.
- Martinec, R. and Salway, A. (2005). A system for image-text relations in new (and old) media. *Visual Communication* 4(3): 337–371.
- Molek-Kozakowska, K. (2014). Hybrid styles in popular reporting on science: A study of new scientist's headlines. In B. Russels, C. Burnier, D. Domingo, F. Le Cam and V. Wiard (eds.), *Hybridity and the news hybrid forms of journalism in the 21st century: Electronic proceedings* (pp. 135–159). Brussels: Vrije Universiteit.
- O'Halloran, K.L. (ed.). (2004). *Multimodal discourse analysis*. London: Continuum.
- O'Halloran, K.L. (2005). *Mathematical discourse: Language, symbolism and visual images*. London: Continuum.

- O'Halloran, K.L. (2011). Multimodal discourse analysis. In K. Hyland and B. Paltridge (eds.), *The continuum companion to discourse analysis*. London: Continuum, pp. 120–137.
- O'Halloran, K.L. and Smith, B.A. (2012a). Multimodal text analysis. In C.A. Chapelle (ed.), *The encyclopedia of applied linguistics*. Hoboken, MA and Oxford: Blackwell Publishing Ltd.
- O'Halloran, K.L. and Smith, B.A. (2012b). Multimodality and technology. In C. Chappell (ed.), *The encyclopedia of applied linguistics*. Hoboken, MA and Oxford: Blackwell Publishers.
- O'Toole, M. (1994/2010). *The language of displayed art*. London and New York: Routledge.
- Reah, D. (1998). *The language of newspapers*. London and New York: Routledge.
- Rohlfing, K., Loehr, D., Duncan, S., Brown, A., Franklin, A. and Kimbarra, I. (2006). Comparison of multimodal annotation tools: Workshop report. *Online-Zeitschrift Zur Verbalen Interaktion* 7: 99–123.
- Rosenbloom, P. (2012). Towards a conceptual framework for digital humanities. *Digital Humanities Quarterly* 6(2). Available at: www.digitalhumanities.org/dhq/vol/6/2/000127/000127.html.
- Terras, M. (2016). A decade in digital humanities. *Journal of Siberian Federal University. Humanities & Social Sciences* 7: 1637–1650.
- Van Leeuwen, T. (1999). *Speech, music, sound*. London: Macmillan Press.
- Van Leeuwen, T. (2005a). *Introducing social semiotics*. London and New York: Routledge.
- Van Leeuwen, T. (2005b). Multimodality, genre, and design. In S. Norris and R.H. Jones (eds.), *Discourse in action: Introducing mediated discourse analysis*. London and New York: Routledge, pp. 73–94.
- Van Leeuwen, T. (2015). Multimodality. In D. Tannen, H.E. Hamilton and D. Schiffrin (eds.), *The handbook of discourse analysis*. Hoboken, MA and Oxford: Blackwell Publishers, pp. 447–465.
- Ventola, E. and Moya, J. (eds.) (2009). *The world told and world shown: Multisemiotic issues*. Hampshire: Palgrave Macmillan.
- Walker, S. (2003). Towards a method for describing the visual characteristics of children's readers and information books. In *Proceedings of the Information Design International Conference*. Recife, Brazil: Sociedade Brasileira de Design da Informação.
- Winston, B. (1990). On counting the wrong things. In M. Alvarado and J.O. Thompson (eds.), *The media reader*. London: British Film Institute, pp. 50–64.

Digital pragmatics of English

Irma Taavitsainen and Andreas H. Jucker

Introduction

Pragmatics, in a very general sense, is the study of the use of language (Verschueren 1999: 1) or more precisely ‘the study of speaker meaning as distinct from word or sentence meaning’ (Yule 1996: 133; but see also Levinson 1983: 5–35, which Huang 2007 takes as his point of departure). In other words, pragmatics studies authentic communication and the negotiation of meaning in interaction. Two traditions are often distinguished, that is, theoretical pragmatics, also called Anglo-American pragmatics, and social pragmatics, or Continental European pragmatics (see Huang 2007: 4–5, 2010: 341; Jucker 2012: 501–503). These two traditions have very different perspectives and manifest themselves in different research paradigms. The first is more concerned with speaker meanings, the way implicit meanings are conveyed (e.g. in Gricean pragmatics) and with cognitive aspects of utterance interpretation. Traugott emphasises this tradition in her definition of historical pragmatics as a ‘usage-based approach to language change’ (2004: 538), dealing with ‘non-literal meaning that arises in language use’ (2004: 539). By contrast, the mainly Continental European view of social pragmatics is broader and examines the social and cultural context of utterance interpretation. According to this view, almost any language phenomenon, including silence, can be scrutinised from a pragmatic perspective: ‘the linguistic phenomena to be studied from the point of view of their usage can be situated at any level of structure or may pertain to any type of form-meaning relationship’ (Verschueren 1999: 2).

The term ‘digital humanities’ was first used as part of a book title by Schreibman et al. (2004) and refers to a field of studies that combines computing with the disciplines of the humanities (see, for instance, Kirschenbaum 2010; Terras 2016). The term ‘digital pragmatics’, as we are going to use it in this chapter, describes a specific way of doing pragmatics. It is not a sub-field of pragmatics, but relies on computerised data and, as we will show in the following, applies a variety of empirical methods to investigate language use, and can be used in many different sub-fields of pragmatics. Electronic text corpora are by far the most important data sources for digital pragmatics, and this is particularly true for diachronic and historical studies. For studies of present-day language use, data sources include not only

digital text corpora but also digital archives of speech and film, as much of modern pragmatic research is multimodal.

Digital corpora have a history of some six decades. The earliest text corpora were compiled at the time when the variationist approach was still approaching and the mainstream focus was on Chomskyan native speaker intuitions. They were compiled in the 1960s and 1970s as pioneering collections of English language materials (e.g. Brown Corpus, Lancaster-Oslo-Bergen Corpus, London-Lund Corpus). Twenty years later the first significant diachronic corpus (Helsinki Corpus) consisted again of texts in English. Since then compilations of digital databases have multiplied in size and number not only in English but in other languages as well, and at present, comprehensive digital archives are found in most European languages. Corpus compilation activity is going on, mainly along two different lines, as the goal can be either mega-corpora of big data or small, structured and annotated specialised databases, often with the compiler's own research as the motivation (as e.g. in the Sociopragmatic corpus, see later).

In this chapter, we will provide an overview of how digital data are used in pragmatic research in various sub-fields of pragmatics. We will assess different aspects of language use that can be studied with the help of digital data and computer technology in general. We will also include a diachronic perspective covering the past, present and – as far as possible – future. In the next section, we will provide a very brief history of pragmatics and how it has been affected by computer technology. In the third section, we will present some of the key issues of digital pragmatics, in particular the issue of the traceability of pragmatic entities. Pragmatic entities tend to have functional definitions (as, for instance, specific speech acts or types of politeness). This makes it difficult to search for them in text corpora because there are no unique search strings that will retrieve all relevant instances. In the next two sections, we will review some of the recent research in the area of corpus pragmatics and introduce the chief methods that have been employed in the field. We will then provide a case study of the speech act of thanking in the annotated, digital seventeenth-century Socio-pragmatic Corpus. Finally, we will discuss future directions of research of this nature.

Background

Pragmatics emerged as a new field of study in the 1960s and 1970s, and initially it did not use digital methods. In the early days, the field was dominated by philosophers, in particular J. L. Austin, John R. Searle and, somewhat later, H. Paul Grice (Austin 1962; Searle 1969; Grice 1975). They used philosophical methods to dissect the essentials of what it means to perform actions with spoken words and how people are able to systematically read between the lines of what their interlocutors say (see Nerlich and Clarke 1996; Nerlich 2009, 2010; Jucker 2012 on the history of pragmatics). An early ground-breaking textbook (Levinson 1983) helped to shape the field. Levinson included not only philosophical approaches in his outline of pragmatics but also what were essentially ethnographic approaches but applied to pragmatic analysis. Such studies investigated small sets of richly contextualised data, and natural everyday conversations were considered to be the optimal sources for such data. In addition to the language philosophers' intuited data, the main data sources of pragmatics at that time were small sets of carefully transcribed conversations, which were very different from the large corpus collections that were generally devoid of language features that occur in natural conversation.

In the early 1990s, computers started to have a more noticeable impact on linguistics in general. Earlier work in corpus linguistics had been based on the corpora from the 1960s,

but it was in the early 1990s, when desktop computers became more widely available, that corpus linguistics started to become one of the main research tools in linguistics. In these early days of corpus linguistics, pragmatically driven work was rare, but there were notable exceptions. Stenström (1991), for instance, investigated a large range of expletives in the London-Lund Corpus and focused specifically on their interactive potential. Tottie (1991) compared the use of backchannels in short extracts of the Santa Barbara Corpus of American English and the London-Lund Corpus containing British English. Stenström and Andersen (1996) investigated the discourse items *cos* and *innit* in the Bergen Corpus of London Teenage Language (COLT), and Aijmer (1996) traced the conversational routines of thanking, apologising, requesting and offering.

Studies in the history of English have always had a somewhat different status, as they cannot rely on native speaker intuition. The only data available are written texts from the period under consideration. As a result, historical linguists were quick to embrace corpus techniques for their studies because these allowed large-scale investigations of actually attested data. The publication of the Helsinki Corpus in 1991 had a major impact on historical linguistics and the study of the history of English. It contains text samples from a wide range of literary and non-literary genres from the earliest extant texts written in English to the end of the early modern period at the beginning of the eighteenth century.

The lack of spoken data made historical research uninteresting for pragmaticists, and historical linguists did not immediately see the potential for pragmatic research questions. It was in the 1990s that the first important collection of papers in historical pragmatics appeared (Jucker 1995), and it already featured some work carried out with corpus methods (Taavitsainen 1995; Nevalainen and Raumolin-Brunberg 1995). Brinton (1996) provided the first large-scale monograph in the area of historical pragmatics investigating grammaticalisation patterns and discourse functions of a range of pragmatic markers in the history of English, from Old English to present-day English. Historical pragmatics developed very quickly and seems to have been one of the important driving forces in the application of corpus methods to pragmatics (for the developments, see Taavitsainen and Jucker 2015).

In the first decade of the twenty-first century, computer technologies and their availability for individual scholars proliferated on an unprecedented scale. Pragmaticists began to take advantage of the potential of large-scale corpora and to develop tools to tackle more complex research questions. Earlier work had relied on relatively simple searches of discourse items (expletives, interjections, backchannels, pragmatic markers and so forth). More recently, scholars have started to extend their searches to speech acts and other more elusive pragmatic entities, and important collected volumes of pragmatic corpus linguistic papers have appeared in a field that is now explicitly called corpus pragmatics. Jucker et al. (2009) consists of the papers that were presented at the first ICAME conference devoted explicitly to corpus-based work in the areas of pragmatics and discourse. Taavitsainen et al. (2014) present a range of corpus pragmatic studies that deal with historical data. In the introduction to the volume, Jucker and Taavitsainen (2014a) outline their understanding of diachronic corpus pragmatics as the intersection of the three fields of historical pragmatics, historical linguistics and corpus linguistics. The recent publication of a handbook in the field of corpus pragmatics (Aijmer and Rühlemann 2015) attests to the current status of the field. An introduction and 16 individual contributions provide a comprehensive overview of the extant corpus-based work on such pragmatic entities as speech acts, pragmatic principles, pragmatic markers, evaluation, reference and turn-taking.

Critical issues and topics

The key issue of digital pragmatics concerns the traceability of pragmatic entities. Corpus linguistic tools of data retrieval depend on explicit search strings. It is easy, for instance, to search for a particular word form. It is also possible to search for syntactic configurations if the texts in the corpus are syntactically annotated. Computers are highly efficient in identifying search strings in text corpora whatever the size of the corpus, but they can only identify elements that can be represented by specific strings. Pragmatic elements, however, are often elements that are defined by their function not by their form. In fact, it has now become standard to distinguish between two different directions of analysis: form-to-function and function-to-form (see Jacobs and Jucker 1995 for what appears to be the original suggestion of this distinction). In a form-to-function approach the researcher takes a linguistic expression as a point of departure, a discourse marker such as *well* or *you know*, for instance, or an address term, such as *you* or *thou*. These forms may co-exist in different spellings (e.g. *y'know*) and undergo slight changes in the course of history, but basically the researcher starts from a specific, clearly identifiable form or a small set of identifiable, related forms and tries to establish the different ways in which this form or these forms are being used or have been used in the course of time. By contrast, a function-to-form approach starts out with a specific pragmatic function. This can be a specific speech act or even more abstract concepts such as evaluation or politeness (see later). The function constitutes whatever a linguistic element achieves in a particular situation, and an analytical category may be either proposed by researchers, such as implicatures, or be regularly used by the members of a given speech community, with speech acts or genre labels providing pertinent examples.

Theoretical philosophical considerations single out aspects that need attention in empirical studies and they serve as analytical categories, but there are severe problems when theories are ‘translated’ into empirical studies on natural data. This is partly due to the fact that speech acts are fuzzy concepts. A specific speech act – a compliment, for instance – may be realised in a shape that is very typical for compliments (i.e. in a prototypical format), or it may be realised in an unusual and less typical shape. Such a prototype approach allows for overlaps and double or even multiple memberships (Fowler 1982), and speech acts can be assessed according to their degree of prototypicality, along several parameters that vary according to the speech act under scrutiny. Prototypical compliments contain positive evaluation and can be distinguished from ironic/sarcastic and insincere compliments which have a negatively evaluative intent. The following turns show the reaction of the hearer/receiver and whether the negative intent is transparent to the target.

(1) <i>AL ROKER:</i>	I like your glasses.
<i>WOMAN-1:</i>	Thank you. (COCA, 2017, NBC Today Show)
(2) <i>WILLIE-GEIST:</i>	So, you know what, that is so ridiculous.
<i>NATALIE-MORALES:</i>	Wow.
<i>WILLIE-GEIST:</i>	Take the day off. Well done.
<i>AL ROKER:</i>	You know what, you are so dumb, you don't even – she shouldn't even come in. (COCA 2013, NBC Today Show)

In (1), the first speaker pays a compliment which is acknowledged immediately by the second speaker. It has a prototypical format (see Manes and Wolfson 1981) and appears to be positive and sincere. In (2), on the other hand, the phrase ‘well done’, which looks like a

compliment, appears to have been uttered with an ironic and perhaps even sarcastic intent, which is made clear by the following speaker's reaction.

Politeness is an important dimension in speech act studies. Speech acts that ask the addressee to do something, so-called directives, as well as speech acts that commit the speaker to do something him- or herself, so-called commissives, constitute face threats in the sense of Brown and Levinson (1987). Directives threaten the negative face of the addressee, to quote Spencer-Oatey (2000: 17) they 'can easily threaten rapport, because they affect our autonomy, freedom of choice, and freedom from imposition, and thus threaten our sense of equity rights (our entitlement to considerate treatment)'. The force of asking can range from a well-meant piece of advice, which leaves the addressee a lot of freedom to comply or not, to commands, which demand immediate compliance and, thus, constitute a serious imposition on the addressee's wish to be unimpeded. In commissives, speakers commit themselves to a specific course of action and thereby threaten their own negative face (i.e. they limit their own freedom). In apologies, to give an example of a speech act that expresses the speaker's feeling, a so-called expressive, the speaker's own positive face is threatened: the speaker acknowledges and takes responsibility for an offence or a potential offence.

Searle (1979: 12–16) proposed several important aspects of speech act theory. The first is the distinction between the illocutionary force of an utterance and its perlocutionary effect. The illocutionary force of an utterance describes its intended contextual meaning. In most speech acts, it is the illocution that counts, for example, according to the definition, a directive speech act depends on the speaker's intention to get the hearer to do something. The perlocutionary effect, on the other hand, describes the actual effect that an utterance has on the hearer, irrespective of whether this effect was intended or not. Utterances can be heard as insults or compliments, for instance, as demeaning or friendly by the hearer, even if the speaker did not intend them as such.

The direction of fit between words and the world is the second point. In directives and commissives, the direction of fit is word-to-world: the speaker wishes or intends something to happen in the world and expresses it in words. With a commissive speech act, the speaker promises to bring about the fit, while in directives it is the addressee who is expected to bring about the fit. The force can vary from a friendly comment to a strict command or a binding oath. The third aspect, the sincerity conditions, are important in a different way, but they are not necessarily found in all speech acts. They stand in opposition to truth conditions and assess not whether an utterance is true or false but whether or not an utterance has been made in good faith or sincerely. Promises, for example, require the speaker's intention to definitely carry out what is being promised in order for them to be made in good faith. Greetings, on the other hand, generally lack sincerity conditions. It does not seem to matter whether they are uttered sincerely or insincerely (Searle 1969: 67). Compliments again depend on sincerity conditions and on the speaker's positive opinion of the entity that is the object of the compliment, as their meanings can be entirely reversed in the case of irony or banter. In extract (3), the speakers explicitly negotiate the illocutionary force of the first speaker's utterance as a compliment. The second speaker queries its status, and the first speaker goes on to confirm that it was meant as a compliment.

- (3) *RICHARDS:* (. . .) I think that the talent that George Bush has – and I say this with real respect – is that rather than tell you the intricacies of what he knows or what he intends to do, he is very good at saying things that are rather all-encompassing. You know, if you – if you said to George, "What time is it?" he would say, we must teach our children to read.

- KING:* Do you mean that as a compliment?
RICHARDS: I am saying it as a compliment. He is very disciplined, and he knows what his message is (COCA, 1999, CNN-King)

Speech acts show both diachronic and synchronic variation according to various sociolinguistic parameters such as gender, age or educational background, but variation according to pragmatic contextual parameters such as situational factors also occur. In written data, especially in literary materials, genre has proved important for speech act conventions, and insults, for instance, have acquired new functions in saints' lives as turning points of the narrative (we will illustrate these points later).

Current contributions and research

The broad 'perspective view' of pragmatics states that any language phenomenon can be considered from a pragmatic angle (see earlier). We will examine this statement next. To begin with the level of phonology, studies on meaning-making intonation patterns within the pragmatic frame have been conducted by Wichmann on 'please' (2004, 2005). She demonstrates how special meanings are generated in interaction by shifting the accent and changing the word order. 'Please' is a routine formula that regularly occurs in requests and therefore counts as an illocutionary force indicating device of a request. But its force may vary, and it can easily turn into an emotional appeal. The sense of urgency can be intensified by changing its position within the utterance and by shifting the accent (Wichmann 2005: 236). In speaker-hearer-oriented uses, *please* serves as 'a courtesy formula which acknowledges debt with greater or lesser sense of obligation' (Wichmann 2004: 1547). Power relations and the dichotomy between 'public' and 'private' also come into play. Hearer-oriented *please* carries a greater sense of obligation and is typical of private speech with a symmetrical speaker-hearer relationship. The rising intonation has a mitigating function in these situations. By contrast, the speaker-based *please* in public contexts may sound like a command when uttered with falling intonation and used in an asymmetrical power situation from the more powerful to the less powerful participant.

In some languages spelling can be used expressively to imitate a particular colloquial tone of voice emphasising the urgency of the plea; this brings another level of language use to focus. The example comes from Finnish, where the English loanword 'please' was spelled with six i's in BASIC INCOME PLIIIIIS in a student demonstration for a basic income for all citizens in Helsinki (March 18, 2010; see Taavitsainen and Pahta 2012).

Specific lexical items or short phrases are relatively easy to trace with corpus linguistic tools. Fetzer (2009), for instance, traces the two hedges *sort of* and *kind of* in a small corpus of political speeches and political interviews in order to account in detail for their collocational patterns, their pragmatic functions and their distribution in monologic and dialogic political discourse. Aijmer (2009) compares the use of *I don't know* and *dunno* in a corpus of spoken texts by Swedish learners of English, and similar data in a corpus of spoken English by native speakers, and Claridge and Kytö (2014) analyse the English degree modifiers *pretty* and *a bit* in a 50-million-word sub-corpus of the Old Bailey Corpus, a collection of court proceedings of over 200,000 trials. They trace the semantic and pragmatic development of these modifiers diachronically from the 1730s to the 1830s. They show that in their data both expressions are in a grammaticalisation process from their use as an adjective/adverb or a noun phrase, respectively, to a degree modifier ('pretty good', 'wait a bit').

Speech acts bring us to the utterance level or even to the discourse level (see later). Adolphs (2008), for instance, studies pragmatic functions in spoken discourse with corpus linguistic tools and takes the speech acts of suggestion and advice as examples. She uses several spoken corpora including the 5-million-word Cambridge and Nottingham Corpus of Discourse in English (CANCODE) to trace speech act expressions, such as *why don't you*, *why don't we*, *why not* and *suggest*, which are used to perform these speech acts.

On a larger scale, corpus linguistic tools have been used to investigate pragmatic phenomena that cannot easily be linked to individual utterances, such as politeness, impoliteness or verbal aggression. Archer (2014), for instance, explores pragmatic phenomena relating to verbal aggression in the Old Bailey Corpus by using semantic-field tags to capture words, phrases and multi-word units related to what she calls an 'aggression space', and Culpeper (2009) uses the 2-billion-word Oxford English Corpus to trace the metalanguage that people use to talk about impoliteness.

Pragmatic analysis always takes the context of an utterance into account, and thus qualitative assessments are needed even in corpus studies that rely on more elaborate search algorithms. The notion of context has received a great deal of attention in recent discussions. Earlier work tended to focus on invariant forms associated with specific speech acts, such as *please*, *pardon* or *excuse* in the case of apologies (Deutschmann 2003). More recent work has adopted a more dynamic view of language use based on the insight that each new turn in communication creates a new context not only in conversation but in written texts as well. Speaker intentions and audience reactions are also considered. This insight can be seen in recent speech act studies on both present-day and historical data, as researchers have realised that second turns of speech acts need to be considered as well for retrieving underlying implications. Responses to expressive speech acts, for instance, range from thanks and paying in kind (returning the insult or the compliment), to playing it down (*not at all*, *don't mention it*) or accepting (*you are welcome*), or silence may follow (gestures, expressions may be described).

Main research methods

Speech act analysis using corpus methods has received a great deal of attention in both present-day and historical studies. Its research methodology is still under development, but results have already been achieved, and some earlier assumptions based on qualitative studies have been revised.

Empirical speech act investigations depend on corpus-based methodologies both for present-day and historical investigations. The problems begin with data retrieval, as it is difficult to find reliable ways of locating non-routine speech act manifestations. Some progress has been made, but appropriate search algorithms (i.e. precise specifications of search strings) are necessary for more comprehensive and systematic surveys. Digital data sources have grown in number and size in recent years, and methods to retrieve data to answer pragmatic research questions have also improved (see Taavitsainen 2015). Big data includes the constantly growing monitor corpora of digital newspapers, as for instance the News on the Web Corpus (NOW), which, at the beginning 2018, contained some 5.69 billion words and continues to grow on a daily basis. Historical resources include the Early English Books Online (EEBO), which contains 755 million words. These and other mega-corpora can be accessed through a website created and maintained by Mark Davies at Brigham Young University (<https://www.english-corpora.org>). The pioneering Helsinki Corpus (mentioned earlier), which was first released in 1991, contained in comparison a very modest 2 million words.

Expressives are perhaps the most difficult and most elusive of the Searlean categories. They have a vast range of different linguistic realisations, from formulaic utterances to creative and unpredictable ones where not only illocutions but perlocutions count. Compliments, insults, thanks and apologies are very different from one another, and a careful study reveals differences to present-day manifestations between past and present ways of performing them (cf. Jucker and Taavitsainen 2014b). Some expressive speech acts can be performed with a limited number of routine utterances and detected relatively easily with current corpus methodologies, whereas others are less formulaic and more creative, and thus more difficult to identify. Some advances have recently been made with compliment formulae that have been extracted from modern material, but historical data pose different challenges to corpus methodologies that need to be developed.

For example, a corpus study to test and develop corpus methodology for retrieving compliments was conducted by Jucker et al. (2008) in order to improve the methodology for future application to historical studies. Manes and Wolfson (1981: 115) collected American compliments by the diary method of participant observation, and according to their observation these compliments display ‘almost total lack of originality’. Holmes (1988: 480) confirmed their result by stating that ‘compliments are remarkably formulaic’. The test was conducted by translating typical patterns proposed by Manes and Wolfson into search strings of syntactic query language and applying the formulae to the 100-million-word BNC. Problems of both precision and recall were encountered, as several examples had the appropriate structure but not the appropriate function. In the next phase, two independent raters analysed the examples manually. This test confirmed the statement that speech acts are fuzzy notions with a subjective element. The conclusion was that contextual qualitative analysis is needed for distinguishing subtle shades of meaning, whether sincere or ironic, in examples with identical wordings. Manual annotation was suggested as a solution, but for practical reasons, this can only be done on relatively small, representative samples but not comprehensively for large digital corpora.

The analysis of speech acts in different cultures relies on the identification of similar speech functions that provide a *tertium comparationis* and remain constant across space or time. We can view diachronic speech act analysis as a special form of contrastive analysis, but there is a fundamental difference: within one language there is continuity from an earlier to a later stage, from Old English to Middle English, to early and late modern English, and finally, to present-day English. This linear development from older stages of language to more recent phases encompasses cultural changes that are important for contextual assessments, as changes were gradual and took place at different speeds in different layers of society (cf. the dissemination of scientific knowledge). In contrastive analysis proper, the realisations of specific speech acts are compared in disparate contexts without continuity from one stage to another.

In order to retrieve functional elements from large text corpora, specific search techniques for finding expressive speech acts have been developed (for an overview see Jucker and Taavitsainen 2013: 94–96). They can be visualised in three overlapping circles (Figure 7.1).

The first circle indicates typical patterns that can help in data retrieval. Positively connotated adjectives are often present in compliments, and searches for such adjectives are likely to render relevant examples, although they cannot catch all compliments in the corpus. Thus, the recall is limited, and so is the precision as most hits will be false hits from the speech-act point of view. *Could you* is a typical way of phrasing an indirect request. *Please* would be a more reliable candidate, listed in the second circle. The most obvious way of locating speech acts in digital data is provided by specific patterns that are known to be typical for a particular speech act. The inversion of subject and auxiliary is typical of a question, for example.

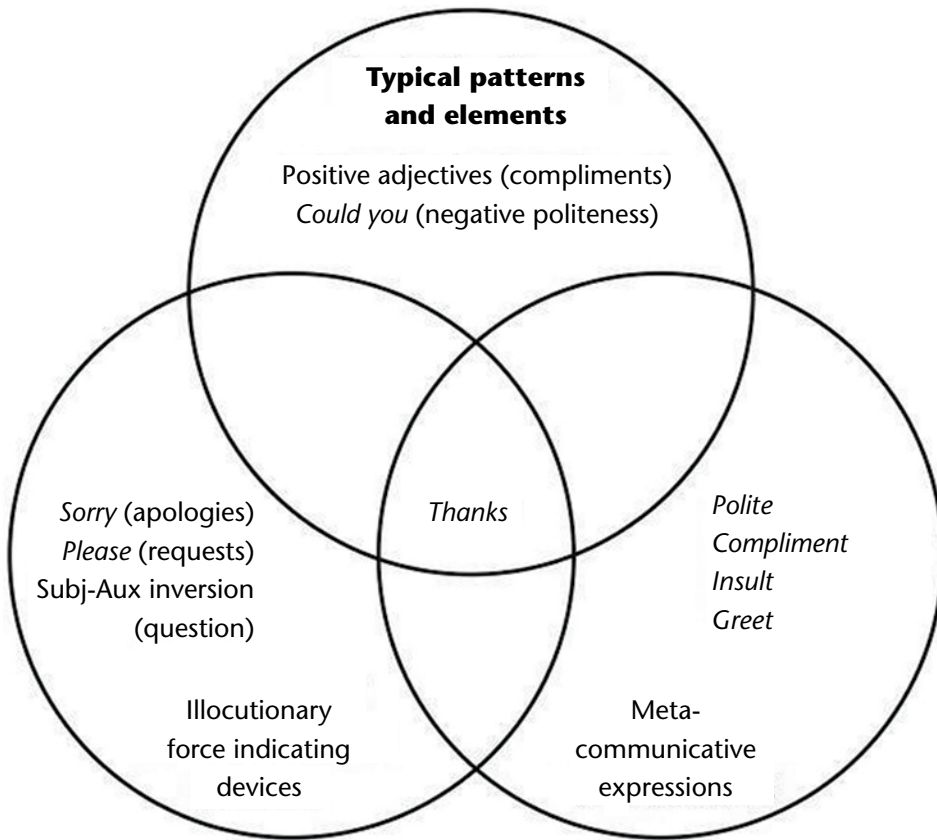


Figure 7.1 Three circles of corpus-based speech act retrieval according to Jucker and Taavitsainen 2013: 95

The second circle gives illocutionary force indicating devices (IFIDs) as search items. Such searches are relatively simple. Deutschmann (2003), for instance, used linguistic elements that typically indicate that an utterance has the illocutionary force of an apology. These were elements such as *sorry*, *pardon*, *excuse* and *apologise*. He took them as reliable indicators of apologies but, of course, it cannot be discounted that the corpus contains some utterances that achieve the function of apologising without using any of his search patterns. There is no guarantee of a perfect recall. There is also the danger that such searches provide many instances that – on closer inspection – turn out not to be apologies. Thus, there is also no guarantee for perfect precision, except that it is usually easier to remove unwanted hits from a retrieved set (in order to improve the precision) than to find those instances that were not retrieved with the search string (which would improve the recall). The next step in Deutschmann's study was to categorise the apologies according to what the offence was that required an apology (material vs mental offences) and the gravity of the case. Sociolinguistic parameters of the speakers and the recipients were also considered. This study provided a model for our diachronic assessment that, however, shows more variation in apologies (Jucker and Taavitsainen 2008).

Other speech acts may be equally ritualised and, therefore, relatively easy to retrieve (e.g. thanking, requests and offers) (Aijmer 1996). But many speech acts are more creative and cannot be identified via IFIDs. Compliments are a case in point. They have been claimed to be surprisingly stereotypical (as pointed out earlier), but there are no IFIDs that would retrieve a significant number of them from a corpus. In historical data, for instance, it turned out that a search for positive adjectives, in spite of both a very low precision and – presumably – a very low recall, provided an interesting and useful collection of compliments (Taavitsainen and Jucker 2008).

The third circle brings us to a different approach. An alternative way of retrieving speech acts in digital data is provided by meta-communicative expression analysis. In an earlier study, we assessed the semantic field of verbal aggression with this method by searching for expressions referring to the speech act by their labels. With this method, we were able to locate passages in which speakers talk about specific speech acts, even though the exact wording is missing, with a few exceptions (Taavitsainen and Jucker 2007). In a subsequent study, Jucker and Taavitsainen (2014b) located compliments by meta-communicative interactions about compliments. They did this by using the search term ‘compliment’ and then located the specific compliment that the interlocutors were talking about at this point.

All these methods are no more than approximations. None can claim to retrieve all instances of a specific speech function, which may or may not be a problem for individual research questions. In the case of searches for meta-communicative expressions this is particularly clear. A search for the term ‘compliment’ will not even come close to identifying all possible compliments in a text because most compliments are paid and received without any explicit mention of the word. But the search retrieves a large number of instances in which the interlocutors found it necessary – for one reason or another – to talk about the status of an utterance as a compliment. They may ask whether an utterance was really intended as a compliment, they may indicate that they take a specific act or utterance as a compliment which was not necessarily intended as such and so on. Such instances therefore provide an ethnographic view of interactants’ attitudes towards compliments. They provide analytical opportunities that would not be available by ordinary compliments that are paid and received without any further metacommunication.

The methods described here often depend on the researcher’s familiarity with a specific speech act and his or her ability to compile an inventory of typical patterns that can be used as search strings. However, researcher intuitions are generally an unreliable guide for a really comprehensive inventory. This is particularly true for historical periods of a language. Kohnen (2007) recommends analysing a sufficiently large but still manageable representative sample of the texts manually in order to identify possible manifestations; this inventory can then be used to retrieve the speech act from a digital database.

Corpus annotation brings speech act studies to a new level as it enables one to focus on sociopragmatic aspects by taking the underlying sociolinguistic speaker parameters into account. The disadvantage is that such annotation requires a great deal of manual work. Therefore, it has been applied to a small set of historical texts only (see the case study next; for a recent overview see Weisser 2015).

Sample analysis

This section includes a case study to demonstrate the methodology of digital corpus pragmatics on seventeenth-century dialogic data.¹ It focuses on speech acts expressing gratitude

and combines with variational pragmatics and contextual analysis. The speech act of thanking has been defined by Searle (1969: 63) by the following set of rules:

- Propositional content rule: past act A done by H (hearer).
- Preparatory rule: A benefits S (speaker) and S believes A benefits S.
- Sincerity rule: S feels grateful or appreciative for A.
- Essential rule: Counts as an expression of gratitude or appreciation.

These rules are broken in ironic uses. Thanking has also gained a discourse function of closing a conversation, and it can be used as a small supportive ritual of polite interaction. *Thank you* and *thanks* are frequent as responses to accepting/rejecting an offer or as acknowledgements of a benefit. Thanking has been previously dealt with from the present-day angle by Aijmer (1996), Schauer and Adolphs (2006), Wong (2010) and Jautz (2013), and historically by Jacobsson (2002).

Thanking is an inherently polite speech act that coincides with a convivial intrinsically courteous or polite function (Leech 1983: 104–105, 2007). An analytic view of thanking pays attention to various aspects that underlie or constitute the speech act. What is thanked for are usually gifts or favours, services, compliments or well wishes, and their positive politeness can be intensified by adverbs, prosodic devices or repetition (Aijmer 1996: 35, Jacobsson 2002: 68). Even address terms have an intensifying function when combined with thanking (Taavitsainen and Jucker 2016). Gratitude expressions are often followed by a second part, and the most usual response is either acceptance or dismissal, with mostly formulaic verbal expressions such as *don't mention it* (British English) and *you're welcome* (American English; see also Rüegg 2014).

The *Oxford English Dictionary* (OED) with its *Historical Thesaurus of English* (HTE) provides definitions that can serve as a point of departure for empirical pragmatic studies.

OED: *Thank you:* a courteous acknowledgement of a favour or service to add emphasis . . . usually implying refusal or dismissal

No, thank you, . . . to imply self-satisfaction *Thanks:* A much abbreviated expression of gratitude *Obliged:* Bound by law, duty, or any moral tie, esp. one of gratitude; now *arch.* and *rare*

HTE: Gratitude 02.04.18 n Up the hierarchy to: The mind; Emotion aj (Grateful), av (Gratefully), n (Thanks), p (Thanks to) ph (Expressions of thanks), v (Give thanks) vi (Give thanks), vt (Thank)

The search items listed here can be complemented by others gleaned from previous studies, for example, in this case they include *thank you*, *thanks*, *thankful*, *grateful*, *appreciate it* and *obliged*. Such ready-made lists are used in top-down methods, and we can assume that most gratitude expressions are found by this method. However, to ensure the coverage, the material was read through in this case for hidden manifestations, but nothing new came up in a qualitative reading.

The present study focuses on social variation in The Sociopragmatic Corpus (1640–1760), which is based on a part of the Corpus of English Dialogues (CED, 1560–1760), but makes use of advanced sociopragmatic annotation that records speaker/hearer relationships, social roles and sociological characteristics such as gender. Its language samples are set in various context(s), which are understood as dynamic (see Archer 2005: chapter 4). In this case, part of the data is fictional: the data reflect real language use, but condense and typify the features, and genre constraints need to be taken into account (see e.g. Fludernik 1993; Taavitsainen 1997).

The aim of the study was to find out how much social variation there is in the performative ways of thanking, whether *obliged* was widely used as a gratitude expression in these seventeenth-century drama texts and whether sociolinguistic speaker parameters reveal any differences. A further set of research questions concerned the development from exclamations to mild swearing and whether any discourse functions of *thanks/thank you* can be found. Such uses come about by the process of pragmaticalisation and include a meaning change whereby a word or a phrase changes its propositional meaning to a pragmatic interpersonal meaning between the persons participating in the interaction (Frank-Job 2006: 360); it turned out that the present material provides an excellent early example of this new meaning.

In a previous study, Jacobsson (2002) used CED materials from several genres and achieved some pertinent results. He found 249 occurrences of thankings: language teaching handbooks had by far the highest number (29 occurrences), as these dialogues were intended for learners of English to set a model for polite language use. Single lexical items of thanking were most common, but he was also able to establish some patterns with the most common collocations: SIR, I/WE THANK YOU (14), I/WE HUMBLY THANK (11), I THANK YOU HARTILY/HEARTILY (4) and I HARTILY THANK (3).

The absolute occurrences of the search items were the following:

Thank (19): *I/ we thank you, how shall I thank or praise thee, I thank God, Thank Heaven, I thank my stars, I thank my fortune, Thank my Wit, thank him*

Thanks (14): *I give you (a thousand/ many) thanks, Thanks be to God, my thanks to you, Present my thanks and best respects to, You shall have my thanks.*

Grateful (once)

Obliged (4 times): *I am very much obliged to you, I am Grateful where I am obliged.*

A closer look at the sociolinguistic parameters of the annotation shed some additional light to language use:

NOBILITY, social elite Example 1: Speaker: female, member of nobility. Addressee: male, member of nobility.

Bev. jun. But Madam, we grow grave methinks – Let's find some other Subject – Pray how did you like the Opera last Night?

Ind. First give me leave to thank you, for my Tickets.

Bev. jun. O! your Servant, Madam – But pray tell me . . . (D5CSTEEL)

Analysis of *thank you*: The phrase is used as an acknowledgement of a gift. The response begins with an exclamation and continues with a self-effacing phrase accompanied by an address term. The tone is refined, with the decorum of courtly language use.

Example 2: Speaker: male, member of nobility. Addressee: male, member of nobility.

Dor. Nay, if he wo't not, I am ready to wait upon the # Ladies; and I think I am the fitter Man.

Sr. Jas. You, Sir, no I thank you for that – Master *Horner* is a privileg'd Man amongst the virtuous Ladies, 'twill be a great while before you are so . . . , Sir, for as I take it, the virtuous Ladies have no business with you. (D3CWYCHE)

Analysis of *thank you*: The phrase is part of a negative reply, adding emphasis to a dis-preferred response to an offer. The turn begins with the second-person pronoun of direct address with politeness and continues by giving reasons for playing the offer down.

Example 3: Speaker: male, member of nobility. Addressee: female, member of nobility.

Lur. D'you think so; thank God I'm as healthy as a Horse; pardon the Comparison.
Flor If the Horse does, I do. . .

Analysis of *thank God*: This is an exclamation of mild swearing, followed by an apology for crude language use. A witty answer in the jocular mode follows.

Example 4: Speaker: male, member of nobility. Addressee: female, member of nobility.

VWil Give me this Night to think in, I'll promise nothing but this: I'm Grateful where I am obliged. (D4MANLE)

Analysis of *I'm Grateful where I am obliged*: It is a tautological phrase used here in polite interaction about marriage negotiations. Tautologies have received various pragmatic analyses (see Ward and Hirschberg 1991), but the jocular attitude found here is not mentioned. The tone is positive, aimed at creating suspense.

Example 5: Speaker: male, member of nobility. The previous turn is very long and contains a curse.

I'm very much obliged to you, Sir, for your kind wishes (D5CMILLE)

Analysis of *obliged*: This is an example of ironical/ sarcastic use as a response to an itinerant physician who throws a curse of all kinds of evils to him. The adjective *kind* and the opposite term *wish* continue in the same vein, enforcing the ironical effect.

GENTRY Example 1: Speaker: female, member of the gentry. Addressee: male, member of the gentry.

Sr. Jas. . . . I have provided an innocent Play-fellow for you there.
Dain. Who he!
Squeam. There's a Play-fellow indeed.
Sr. Jas. Yes sure, what he is good enough to play at # Cards, . . .
Squeam. Foh, we'll have no such Play-fellows.
Dain. No, Sir, you shan't choose Play-fellows for us, we thank you.
Sr. Jas. Nay, pray hear me. (*Whispering to them.*)
Lad. But, poor Gentleman, cou'd you be so generous? . . . (D3CWYCHE)

Analysis of *we thank you*. The phrase is used to emphasise a negative response with a clear indication that the discussion on the topic is finished. The pronoun *we* adds to its force. No gratitude is involved, but thanking has a clear discourse function expressing an indirect interpersonal meaning beyond what is said; the response shows that the hearer understood the implication but does not want to comply.

In an article about insults (Jucker and Taavitsainen 2000), we developed the notion of pragmatic space and claimed that manifestations of speech acts can best be studied in a multidimensional pragmatic space with several continua that can vary according to the speech act in question. As mentioned earlier, speech acts are fuzzy concepts, and this scheme is helpful with thanking as well. Especially the dimensions of speaker attitudes and responses seem important. The material contains sincere gratitude expressions, but irony and jocularity are also present. Responses to thanking vary from neutral acceptance to a humble response, reflecting politeness norms of the social elite. *Obliged* seems to be used by the highest classes (the jocular example is from 1686, see also Taavitsainen and Jucker 2010). Thus, some social variation was attested in this pilot study, but more data are needed for more precise conclusions.

Future directions

The snapshots provided in this chapter give us glimpses of pragmatic developments in the history of English with digital searches of small-scale corpora. The method applied in most empirical corpus studies on speech acts has been named ‘illustrative eclecticism’ (Kohnen 2004: 241), which describes the present state of the art in an adequate way. Computerised searches of large corpora are very different and need appropriate search algorithms. Lutzky and Kehoe (2017), for instance, want to trace apology expressions in the 600-million-word Birmingham Blog Corpus, which makes it impossible to manually differentiate between expressions such as *sorry* that are used in an apology from their homonymous uses outside of apologies (e.g. ‘a sorry state of affairs’). Therefore, they establish collocational profiles for each apology expression and use the lists of retrieved collocates to distinguish between different uses of their search terms and to identify further apology expressions in their data.

At present, corpora are being developed into two somewhat opposite directions. Modern spoken or multimodal pragmatic corpora and historical philologically oriented corpora both provide contextual anchoring that enable micro-level assessments. Annotated small corpora are well suited for corpus pragmatic research tasks, as the previous case study demonstrates, but the downside is that they remain necessarily small in size, as they require a great deal of qualitative analysis in preparation. This fact lessens the validity of the results. The opposite trend is towards ever larger digital databases. Contextualisation, however, is difficult, and methodological innovations are needed for corpus pragmatic research. A recent socio-pragmatic study by McEnery and Baker (2016) uses the huge Early English Books Online database for lexical searches and applies qualitative analyses to the relevant passages. The results are then related to the seventeenth-century sociocultural developments. This model could be applied to other topics as well. It will be exciting to see how the field develops in the future.

Note

- 1 This case study is based on a talk at the IPrA Conference 2015 by Taavitsainen. I am grateful to Dawn Archer for letting me use the Sociopragmatic Corpus for this study.

Further reading

- 1 Aijmer, K. (1996). *Conversational routines in English: Convention and creativity*. London: Longman.

This is one of the early studies using corpus methodologies in order to study pragmatic entities. It focuses on what it calls conversational routines, that is to say on speech acts that tend to be realised in routinised forms (i.e. thanking, apologising, requesting and offering).

- 2 Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.
This textbook provides a survey of corpus linguistic tools that can be used for the analysis of natural discourse. Introductions are provided to issues of frequency and dispersion, concordances, collocations and keyness.
- 3 Adolphs, S. (2008). *Corpus and context: Investigating pragmatic functions in spoken discourse*. Amsterdam and Philadelphia: John Benjamins.
This monograph explores the potential of corpus methodology in the analysis of speech functions. The speech act of suggesting or advice giving is taken as an example, and its contextual occurrences are analysed in a number of large spoken corpora.
- 4 Taavitsainen, I., Jucker, A.H. and Tuominen, J. (eds.) (2014). *Diachronic corpus pragmatics*. Amsterdam: John Benjamins.
This volume of articles contains work that focuses on diachronic developments of specific language uses investigated on the basis of large text corpora. The majority of papers are devoted to the history of English, but other languages are included as well (Finnish, Japanese, Estonian, Dutch, Swedish, Old Spanish).
- 5 Aijmer, K. and Rühlemann, C. (eds.) (2015). *Corpus pragmatics: A handbook*. Cambridge: Cambridge University Press.
This handbook provides state-of-the-art accounts of a broad range of sub-fields in pragmatics that use corpus linguistic methodologies. These sub-fields include speech acts, pragmatic principles, pragmatic markers, evaluation, reference and turn-taking. It also contains a programmatic introduction that provides a useful survey of corpus pragmatics.

References

- Adolphs, S. (2008). *Corpus and context: Investigating pragmatic functions in spoken discourse*. Amsterdam: John Benjamins.
- Aijmer, K. (1996). *Conversational routines in English. Convention and creativity*. London: Longman.
- Aijmer, K. (2009). 'So er I just sort I dunno I think it's just because . . .': A corpus study of *I don't know* and *dunno* in learners' spoken English. In A.H. Jucker, D. Schreier and M. Hundt (eds.), *Corpora: Pragmatics and discourse. Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29). Ascona, Switzerland, 14-18 May 2008* (Language and Computers: Studies in Practical Linguistics 68). Amsterdam: Rodopi pp. 151–168.
- Aijmer, K. and Rühlemann, C. (eds.) (2015). *Corpus pragmatics: A handbook*. Cambridge: Cambridge University Press.
- Archer, D. (2005). *Questions and answers in the English courtroom (1640–1760)*. Amsterdam and Philadelphia: John Benjamins.
- Archer, D. (2014). Exploring verbal aggression in English historical texts using USAS. The possibilities, the problems and potential solutions. In I. Taavitsainen, A.H. Jucker and J. Tuominen (eds.), *Diachronic corpus pragmatics*. Amsterdam: John Benjamins, pp. 277–301.
- Austin, J.L. (1962). *How to do things with words: The William James lectures delivered at Harvard university in 1955*. Oxford: Oxford University Press.
- Brinton, L.J. (1996). *Pragmatic markers in English: Grammaticalization and discourse functions*. Berlin: Mouton de Gruyter.
- Brown, P. and Levinson, S.C. (1987). *Politeness. Some universals in language usage*. Cambridge: Cambridge University Press.

- Claridge, C. and Kytö, M. (2014). 'I had lost sight of them then for *a bit*, but I went on *pretty* fast': Two degree modifiers in the Old Bailey Corpus. In I. Taavitsainen, A.H. Jucker and J. Tuominen (eds.), *Diachronic corpus pragmatics*. Amsterdam: John Benjamins, pp. 29–52.
- Culpeper, J. (2009). The metalanguage of impoliteness: Using sketch engine to explore the Oxford English corpus. In P. Baker (ed.), *Contemporary corpus linguistics*. London: Continuum, pp. 64–86.
- Deutschmann, M. (2003). *Apologising in British English*. Umeå: Institutionen för moderna språk, Umeå University.
- Fetzer, A. (2009). *Sort of* and *kind of* in political discourse: Hedge, head of NP or contextualization cue? In A.H. Jucker, D. Schreier and M. Hundt (eds.), *Corpora: Pragmatics and discourse. Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29). Ascona, Switzerland, 14–18 May 2008* (Language and Computers: Studies in Practical Linguistics 68). Amsterdam: Rodopi, pp. 127–149.
- Fludernik, M. (1993). *The fictions of language and the languages of fiction*. London: Routledge.
- Fowler, A. (1982). *Kinds of literature: An introduction to the theory of genre and modes*. Oxford: Clarendon.
- Frank-Job, B. (2006). A dynamic-interactional approach to discourse markers. In K. Fischer (ed.), *Approaches to discourse particles*. Amsterdam: Elsevier, pp. 359–374.
- Grice, P.H. (1975). Logic and conversation. In P. Cole and J.L. Morgan (eds.), *Syntax and semantics 3: Speech acts*. New York: Academic Press, pp. 41–58.
- Holmes, J. (1988). Paying compliments: A sex preferential politeness strategy. *Journal of Pragmatics* 12(4): 445–465.
- Huang, Y. (2007). *Pragmatics*. Oxford: Oxford University Press.
- Huang, Y. (2010). Pragmatics. In L. Cummings (ed.), *The pragmatics encyclopedia*. London: Routledge, pp. 341–345.
- Jacobs, A. and Jucker, A.H. (1995). The historical perspective in pragmatics. In A.H. Jucker (ed.), *Historical pragmatics. Pragmatic developments in the history of English* (Pragmatics & Beyond New Series 35). Amsterdam/Philadelphia: John Benjamins, pp. 3–33.
- Jacobsson, M. (2002). 'Thank you' and 'thanks' in Early Modern English. *ICAME Journal* 26: 63–80.
- Jautz, S. (2013). *Thanking formulae in English. Explorations across varieties and genres*. Amsterdam: John Benjamins.
- Jucker, A.H. (ed.) (1995). *Historical pragmatics: Pragmatic developments in the history of English*. Amsterdam and Philadelphia: John Benjamins.
- Jucker, A.H. (2012). Pragmatics in the history of linguistic thought. In K. Allan and K. Jaszczołt (eds.), *The Cambridge handbook of pragmatics*. Cambridge: Cambridge University Press, pp. 495–512.
- Jucker, A.H., Schneider, G., Taavitsainen, I. and Breustedt, B. (2008). Fishing for compliments: Precision and recall in corpus-linguistic compliment research. In A.H. Jucker and I. Taavitsainen (eds.), *Speech acts in the history of English* (Pragmatics & Beyond New Series 176). Amsterdam/Philadelphia: John Benjamins, pp. 273–294.
- Jucker, A.H., Schreier, D. and Hundt, M. (eds.) (2009). *Corpora: Pragmatics and discourse: Papers from the 29th international conference on English language research on computerized corpora (ICAME 29). Ascona, Switzerland, 14–18 May 2008*. Amsterdam: Rodopi.
- Jucker, A.H. and Taavitsainen, I. (2000). Diachronic speech act analysis: Insults from flying to flaming. *Journal of Historical Pragmatics* 1(1): 67–95.
- Jucker, A.H. and Taavitsainen, I. (2008). Apologies in the history of English: Routinized and lexicalized expressions of responsibility and regret. In A.H. Jucker and I. Taavitsainen (eds.), *Speech acts in the history of English* (Pragmatics & Beyond New Series 176). Amsterdam/Philadelphia: John Benjamins, pp. 229–244.
- Jucker, A.H. and Taavitsainen, I. (2013). *English historical pragmatics*. Edinburgh: Edinburgh University Press.
- Jucker, A.H. and Taavitsainen, I. (2014a). Diachronic corpus pragmatics: Intersections and interactions. In I. Taavitsainen, A.H. Jucker and J. Tuominen (eds.), *Diachronic corpus pragmatics* (Pragmatics & Beyond New Series 243). Amsterdam: John Benjamins, pp. 3–26.

- Jucker, A.H. and Taavitsainen, I. (2014b). Complimenting in the history of American English: A meta-communicative expression analysis. In I. Taavitsainen, A.H. Jucker and J. Tuominen (eds.), *Diachronic corpus pragmatics* (Pragmatics & Beyond New Series 243). Amsterdam: John Benjamins, pp. 257–276.
- Kirschenbaum, M.G. (2010). What is digital humanities and what's it doing in English departments? *ADE Bulletin* 150: 55–61.
- Kohnen, T. (2004). Methodological problems in corpus-based historical pragmatics: The case of English directives. In K. Aijmer and B. Altenberg (eds.), *Advances in corpus linguistics. Papers from the 23rd international conference on English language research on computerized corpora (ICAME 23) Göteborg 22–26 May 2002*. Amsterdam and New York: Rodopi, pp. 237–247.
- Kohnen, T. (2007). Text types and methodology of diachronic speech act analysis. In S. Fitzmaurice and I. Taavitsainen (eds.), *Methodological issues in historical pragmatics*. Berlin: Mouton de Gruyter, pp. 139–166.
- Leech, G.N. (1983). *Principles of pragmatics*. London: Longman.
- Leech, G. (2007). Politeness: Is there an East-West divide? *Journal of Politeness Research* 3: 167–206.
- Levinson, S.C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.
- Lutzky, U. and Kehoe, A. (2017). 'I apologise for my poor blogging': Searching for apologies in the Birmingham Blog Corpus. *Corpus Pragmatics* 1: 37–56.
- Manes, J. and Wolfson, N. (1981). The compliment formula. In F. Coulmas (ed.), *Conversational routine: Explorations in standardized communication situations and prepatterned speech*. The Hague: Mouton, pp. 115–132.
- McEnery, A. and Baker, H. (2016). *Corpus linguistics and 17th-century prostitution*. London: Bloomsbury.
- Nerlich, B. (2009). History of pragmatics. In J.L. Mey (ed.), *Concise encyclopedia of pragmatics* (2nd ed.). Amsterdam: Elsevier, pp. 328–335.
- Nerlich, B. (2010). History of pragmatics. In L. Cummings (ed.), *The pragmatics encyclopedia*. London: Routledge, pp. 192–195.
- Nerlich, B. and Clarke, D.D. (1996). *Language, action and context. The early history of pragmatics in Europe and America, 1780–1930*. Amsterdam: Benjamins.
- Nevalainen, T. and Raumolin-Brunberg, H. (1995). Constraints on politeness: The pragmatics of address formulae in Early English correspondence. In A.H. Jucker (ed.), *Historical pragmatics: Pragmatic developments in the history of English* (Pragmatics & Beyond New Series 35). Amsterdam/Philadelphia: John Benjamins, pp. 541–601.
- Rüegg, L. (2014). Thanks responses in three socio-economic settings: A variational pragmatics approach. *Journal of Pragmatics* 71: 17–30.
- Schauer, G.A. and Adolphs, S. (2006). Expressions of gratitude in corpus and DCT data: Vocabulary, formulaic sequences, and pedagogy. *System* 34: 119–134.
- Schreibman, S., Siemens, R. and Unsworth, J. (eds.) (2004). *A companion to digital humanities* (Blackwell Companions to Literature and Culture). Oxford: Blackwell.
- Searle, J.R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.
- Searle, J.R. (1979). *Expression and meaning: Studies in the theory of speech acts*. Cambridge: Cambridge University Press.
- Spencer-Oatey, H. (2000). *Culturally speaking: Managing rapport through talk across cultures*. London and New York: Continuum.
- Stenström, A.-B. (1991). Expletives in the London-Lund corpus. In K. Aijmer and B. Altenberg (eds.), *Advances in corpus linguistics. Papers from the 23rd international conference on English language research on computerized corpora (ICAME 23) Göteborg 22–26 May 2002*. Amsterdam and New York: Rodopi, pp. 239–253.
- Stenström, A.-B. and Andersen, G. (1996). More trends in teenage talk: A corpus-based investigation of the discourse items *cos* and *innit*. In C.E. Percy, C.F. Meyer and I. Lancashire (eds.), *Synchronic*

- corpus linguistics: Papers from the sixteenth international conference on English language research on computerized corpora (ICAME 16)*. Amsterdam: Rodopi, pp. 189–203.
- Taavitsainen, I. (1995). Interjections in early modern English: From imitation of spoken to conventions of written language. In A.H. Jucker (ed.), *Historical pragmatics: Pragmatic developments in the history of English* (Pragmatics & Beyond New Series 35). Amsterdam/Philadelphia: John Benjamins, pp. 439–465.
- Taavitsainen, I. (1997). Genre conventions: Personal affect in fiction and non-fiction in early modern English. In M. Rissanen, M. Kytö and K. Heikkonen (eds.), *English in transition: Corpus-based studies in linguistic variation and genre styles*. Berlin and New York: Mouton de Gruyter, pp. 185–266.
- Taavitsainen, I. (2015). Historical pragmatics. In D. Biber and R. Reppen (eds.), *Cambridge handbook on corpus linguistics*. Cambridge: Cambridge University Press, pp. 252–268.
- Taavitsainen, I. and Jucker, A.H. (2007). Speech acts and speech act verbs in the history of English. In S. Fitzmaurice and I. Taavitsainen (eds.), *Methods in historical pragmatics*. Berlin: Mouton de Gruyter, pp. 107–138.
- Taavitsainen, I. and Jucker, A.H. (2008). ‘Methinks you seem more beautiful than ever’: Compliments and gender in the history of English. In A.H. Jucker and I. Taavitsainen (eds.), *Speech acts in the history of English* (Pragmatics & Beyond New Series 176). Amsterdam/Philadelphia: John Benjamins, pp. 195–228.
- Taavitsainen, I. and Jucker, A.H. (2010). Expressive speech acts and politeness in eighteenth-century English. In R. Hickey (ed.), *Eighteenth-Century English: Ideology and change*. Cambridge: Cambridge University Press, pp. 159–181.
- Taavitsainen, I. and Jucker, A.H. (2015). Twenty years of historical pragmatics: Origins, developments and changing thought styles. *Journal of Historical Pragmatics* 16(1): 1–25.
- Taavitsainen, I. and Jucker, A.H. (2016). Forms of address. In C. Hough (ed.), *Oxford handbook of names and naming*. Oxford: Oxford University Press, pp. 427–437.
- Taavitsainen, I., Jucker, A.H. and Tuominen, J. (eds.) (2014). *Diachronic corpus pragmatics*. Amsterdam: John Benjamins.
- Taavitsainen, I. and Pahta, P. (2012). Appropriation of the English politeness marker *Pliis* in Finnish discourse. In L. Frentiu and L. Pugna (eds.), *A journey through knowledge*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 182–203.
- Terras, M. (2016). A decade in digital humanities. *Journal of Siberian Federal University Humanities & Social Sciences* 7: 1637–1650.
- Tottie, G. (1991). Conversational style in British and American English: The case of backchannels. In K. Aijmer and B. Altenberg (eds.), *Advances in corpus linguistics. Papers from the 23rd international conference on English language research on computerized corpora (ICAME 23) Göteborg 22–26 May 2002*. Amsterdam and New York: Rodopi, pp. 254–271.
- Traugott, E.C. (2004). Historical pragmatics. In L.R. Horn and G. Ward (eds.), *The handbook of pragmatics*. Oxford: Blackwell, pp. 538–561.
- Verschueren, J. (1999). *Understanding pragmatics*. London: Arnold.
- Ward, G.L. and Hirschberg, J. (1991). A pragmatic analysis of tautological utterances. *Journal of Pragmatics* 15: 507–520.
- Weisser, M. (2015). Speech act annotation. In Aijmer, K. and Rühlemann, C. (eds.), *Corpus pragmatics: A handbook*. Cambridge: Cambridge University Press, pp. 84–113.
- Wichmann, A. (2004). The intonation of *please*-requests: A corpus-based study. *Journal of Pragmatics* 36: 1521–1549.
- Wichmann, A. (2005). *Please*: From courtesy to appeal: The role of intonation in the expression of attitudinal meanings. *English Language and Linguistics* 9(2): 229–253.
- Wong, M.L.-Y. (2010). Expressions of gratitude by Hong Kong speakers of English: Research from the international corpus of English in Hong Kong (ICE-HK). *Journal of Pragmatics* 42: 1243–1257.
- Yule, G. (1996). *Pragmatics*. Oxford: Oxford University Press.

Metaphor

Wendy Anderson and Elena Semino

Introduction

Metaphor involves communicating and thinking about one thing in terms of another, where the two things are different but some form of similarity can be perceived between them. For example, in the expression ‘the digital turn’, the noun ‘turn’ is used metaphorically: the changes that resulted from the invention and adoption of digital technologies are described in terms of a change in the direction of physical movement. This metaphor relies on a perception of a similarity between changes in physical direction, on the one hand, and abstract and complex societal/cultural changes, on the other.

But why a chapter on *metaphor* in a handbook on English language and the digital humanities? Metaphor is an important topic in the *humanities* for at least two main reasons. First, the ability to talk, perceive and think about one thing in terms of another is an essentially human characteristic, which greatly expands our potential for communication, imagination, reasoning and problem solving. Second, metaphors are one of the main ways in which we express what it is like to be human, whether in literature, song, visual art or everyday informal interactions. Without metaphor, it would be very difficult to talk and think about many of our most intense subjective experiences, such as pain or love, or about some of the big issues of human existence, such as time or death. Digital humanities shares with the broader humanities a concern with understanding the nature of humanity. For Terras (2016: 1637–1638), digital humanities indeed comprises ‘computational methods that are trying to understand what it means to be human, in both our past and present society’.

Over the last 40 years, metaphor has increasingly been seen as a central tool for *both* communication *and* thinking, largely through the influence of conceptual metaphor theory (Lakoff and Johnson 1980) and of subsequent cognitive scientific approaches to thinking such as Fauconnier and Turner’s (2002) theory of conceptual integration. Within conceptual metaphor theory in particular, conventional metaphorical expressions are seen as evidence of conventional patterns of metaphoric thought known as ‘conceptual metaphors’. For example, *turn* in *the digital turn* can be seen as a realisation of the conceptual metaphor CHANGE IS MOTION (Kövecses 2010: 187), whereby the target domain CHANGE is understood in terms of the source domain MOTION.

Metaphor is therefore important in the study of any language, but the *English language* has so far been the focus of most studies, possibly at the expense of other languages. Kirschenbaum (2013: 195) argues that ‘digital humanities has accumulated a robust professional apparatus that is probably more rooted in English than any other departmental home’. Indeed, one of the earlier designations of the field, ‘literary and linguistic computing’, stresses the connections to English, both language and literature. Metaphor also tends to be involved in new conceptual, cultural and communicative phenomena, and the *digital* is no exception. For example, many new concepts, activities and phenomena associated with the digital turn are labelled via metaphorical extensions of the meanings of words that already existed, such as *web*, *surfing*, *mouse*, *bug* and so on. In that way, our knowledge and experiences of the physical world are exploited both to name and understand our experiences of the digital world.

In this chapter, we focus on the intersection of metaphor studies and the digital humanities, beginning by discussing how digital data, methods and tools are increasingly being used in research on metaphor. We then present two case studies from larger projects on, respectively, metaphors for cancer in online interactions among patients and metaphors involving the conceptual domain of REPTILES in a ‘map’ of English metaphors based on the Historical Thesaurus of English. We finish by highlighting some critical issues and suggesting some possible directions for future work.

Current research and methods

In a discussion of the challenges and opportunities of digital methods in the humanities, Rieder and Röhle comment on the ‘explosion of material available in digital form’ (from online versions of printed books to computer games) and add that:

researchers are increasingly turning towards automated methods of analysis in order to explore these artifacts and the human realities they are entwined with. These computational tools hold a lot of promise: they are able to process much larger *corpora* than would ever be possible to do manually; they provide the ability to seamlessly zoom from micro to macro and suggest that we can reconcile breadth and depth of analysis; they can help to reveal patterns and structures which are impossible to discern with the naked eye; some of them are even probing into that most elusive of objects, *meaning*.

(Rieder and Röhle 2012: 67, *italics in original*)

The various lines of research that we look at in this section do indeed probe into that ‘elusive’ phenomenon of ‘meaning’, as metaphors are centrally involved in how meanings are made and evolve in the language system, in conceptual structure and in communication. In showing how Rieder and Röhle’s general considerations apply to research on metaphor here, we discuss:

- 1 The use of corpus linguistic methods to study metaphor in large datasets
- 2 The creation of digital resources and tools in the study of metaphor
- 3 The study of metaphor in digital communication

In all of these cases, digital methods and resources enable us to see patterns that would otherwise be hidden.

Corpus-based approaches to the study of metaphor

Since the mid-2000s, major advances in the study of metaphor have been achieved by the adoption of the methods of corpus linguistics – the computer-aided analysis of large digital collections of texts known as ‘corpora’ (McEnery and Hardie 2012) and one of the areas of digital humanities with the longest history.

The application of corpus methods is particularly appropriate in view of the claims made in conceptual metaphor theory that metaphors are frequent in language and that systematic conventional patterns of linguistic metaphors reflect conventional patterns of metaphoric thinking, known as ‘conceptual metaphors’ (Lakoff and Johnson 1980). In the seminal studies that established conceptual metaphor theory as the dominant paradigm in metaphor research, linguistic evidence for conceptual metaphors consisted of decontextualised sentences drawn from the authors’ intuitions and/or memories about (English) language use (Lakoff and Johnson 1980, 1999; Lakoff 1993). Not surprisingly, therefore, early corpus-based studies of metaphor took a critical stance to the approach to data and evidence by Lakoff and his colleagues and aimed to test the claims of conceptual metaphor theory by analysing metaphor use in the largest available corpora of English.

Deignan (2005) used a 56-million-word cross-section of the Bank of English Corpus to test, challenge and refine the claims made by conceptual metaphor theory. She provided systematic evidence, for example, for conceptual metaphors such as COMPLEX ABSTRACT SYSTEMS ARE PLANTS (e.g. *Laser equipment [. . .] can be used in many branches of surgery*; Kövecses 2010: 127). Deignan also showed, however, that plant-related metaphors tend to be used more frequently for particular types of complex systems (notably businesses, relationships, ideas and people) and pointed out patterns of metaphor use that conceptual metaphor theory cannot predict or explain. For example, she notes that:

words referring to people and businesses are frequently the subject of metaphorical *blossom* and *flourish*, but not *wither*. In contrast, metaphorical *flower* and its inflections are strongly associated with creative projects, and not used to talk about businesses in the corpus.

(Deignan 2005: 176)

In a similar vein, Stefanowitsch (2006) investigated metaphors for emotions in the 100-million-word British National Corpus, and both confirmed and refined earlier work by Kövecses (2000) in conceptual metaphor theory. The British National Corpus is also used in a more recent study by Johansson Falck and Gibbs (2012) to provide evidence for the role of embodied experiences in the development of conventional metaphorical meanings. Pairs of words that have similar literal meanings, such as *road* and *path*, were found to have different metaphorical meanings in the corpus: *road* is typically used to describe purposeful activities (e.g. *the road to riches and power*), while *path* is used to refer to ways of living and is associated with difficulties (e.g. *the path of drinks and drugs*). Johansson Falck and Gibbs explain these differences in use in terms of differences in the mental imagery associated with the literal uses of each word in responses to a questionnaire about people’s experiences with paths and roads. Overall, this line of research contributes both to metaphor theory (particularly with respect to the relationship between metaphor in language and metaphor in thought) and to the study of a particular language (in this case, English), with implications for teaching and dictionary making.

A large number of corpus-based studies of metaphor, in contrast, exploit specialised and often purpose-built corpora to investigate metaphor use in particular (types of) texts, such as news reports (Musolff 2004; Koller 2004), religious texts (Charteris-Black 2004), political party manifestoes and leaders' speeches (L'Hôte 2014) and online cancer patient forums (Demmen et al. 2015). These studies mostly aim to reveal the functions that metaphors perform and their emotive and ideological implications in particular communicative contexts and historical periods. The use of corpus methods in such cases makes it possible to link qualitative and quantitative analyses of metaphor in a principled way. For example, Salahshour (2016) studies the use of liquid metaphors for immigration (e.g. *the net inflow of migrants*) in a corpus of Auckland-based newspapers in New Zealand from 2007 and 2008. By analysing individual uses in context, she shows that, in contrast to what has been found in similar data from other countries, liquid metaphors are often used in her corpus to present immigration as a positive phenomenon, rather than as a problem. This is seen as a reflection of the particular geographical, cultural and historical background reflected in the corpus.

Other corpus-based studies of metaphor in specialised corpora compare the linguistic representation of similar topics in different genres. For example, differences have been found in the use of metaphors in business English textbooks and corpora of specialist and non-specialist business articles (Skorczynska and Deignan 2006; Skorczynska 2010). This is relevant both for a proper understanding of register and genre variation in metaphor use (see also Deignan et al. 2013) and for the selection of teaching materials in courses preparing students for advanced studies in English language educational institutions.

The vast majority of corpus-based studies of metaphor focus on data from a single language, particularly English. However, some studies take a cross-linguistic approach, including a comparison between metaphor use in English and another language. For example, both similarities and differences were found in the use of body-part metaphors in general corpora of English and Italian (Deignan and Potter 2004) and in the use of blood-related metaphors in Hungarian and American English (Simó 2011; see also Charteris-Black 2001). These studies tend to explain their findings as evidence of both the embodied basis of many metaphors, on the one hand, and broad cultural differences, on the other.

Corpus-based approaches to metaphor mostly rely on standard tools to identify potential uses of metaphor in corpora. Concordancing tools make it possible to find all instances of particular words or phrases in context. These tools can be used to find instances of (1) words that are likely to be used metaphorically in a particular corpus (e.g. *journey* in a corpus of online writing about cancer; Demmen et al. 2015); (2) literally used words referring to concepts that are likely to be described metaphorically in the immediate co-text (e.g. words referring to emotions such as *fear*; Stefanowitsch 2006); and (3) words or phrases that often introduce figurative uses of language, known as 'signalling expressions' or 'tuning devices' (e.g. *sort of* and *like*; Cameron and Deignan 2003; Anderson 2013). Several recent studies have used a tool for the semantic annotation of corpus data to obtain semantic concordances, that is, all instances of expressions belonging to semantic fields that are likely to include metaphorical expressions (e.g. the semantic domain of 'Warfare' in a corpus of online writing about cancer; Demmen et al. 2015). Collocation tools, which identify the words that tend to occur in close proximity to other words, have also been used to study metaphors in corpora. Some studies examine the collocates of literal expressions to discover whether and how the corresponding concept is talked about metaphorically (e.g. Deignan's 2005 analysis of the collocates of *price*). Other studies examine the collocates of metaphorical expressions to discover what topics they are applied to and with what implications (e.g. L'Hôte's 2014 analysis of the collocates of *tough* in the UK Labour Party's manifestoes).

In all these cases, however, the output of corpus tools needs to be analysed manually in order to see whether the search items, or words occurring in close proximity to search items, are in fact used metaphorically. The aim to automate metaphor identification completely is, however, being pursued in the areas of computational linguistics and natural language processing, as we explain in the next section.

Digital resources and tools in the study of metaphor

It is fair to say that text has been at the forefront of the emerging discipline of digital humanities, as indeed it had been under earlier configurations such as literary and linguistic computing and humanities computing. This is because, as Kirschenbaum has explained, ‘after numeric input, text has been by far the most tractable data type for computers to manipulate’ (2013: 201). However, the same cannot be said for figurative language, and it has recently been commented that ‘our ability to handle non-literal language in digital humanities is not yet fully formed’ (Alexander and Bramwell 2014). Generally, the presence of the word *metaphor* in the index of a book on digital humanities points not to research into metaphor, but to metaphors used in the construction of disciplinary identity (for a particularly thought-provoking example, see McCarty 2013). This section is therefore intended to give a flavour of the research and related activities that have used digital resources and tools to investigate metaphor.

While we may not yet have fully cracked the nut of figurative language in digital humanities, research into metaphor has taken advantage of the digital environment since the early days of conceptual metaphor theory. The Berkeley Master Metaphor list has been available online as a research tool since the early 1990s, originally on the website of the University of California, Berkeley (now reproduced at www.lang.osaka-u.ac.jp/~sugimoto/MasterMetaphorList/MetaphorHome.html) and also in document form (see Lakoff et al. 1991). It brings together the conceptual metaphors identified over the first ten years of research since Lakoff and Johnson’s foundational *Metaphors We Live By* (1980) and remains a point of reference despite never being intended as definitive and despite its limitations for some types of research, especially on the more computational side of the field (Loenneker-Rodman and Narayanan 2012).

Perhaps inevitably, given the appropriateness of the short-text format, metaphor has found its way onto Twitter. A Master Metaphor List account was active in 2011, tweeting a conceptual metaphor and a linguistic manifestation of that metaphor on a near-daily basis (@MetaphorList). Other Twitter feeds on conceptual metaphor (active at the time of writing) include Metaphor Awareness (@metaphoraware) and Conceptual Metaphors (@women_fire). Tweets by the lexicographers of Oxford Dictionaries (@OxfordWords) and the OED (@OED) often feature metaphor, generally with some insight into their origin. In addition to tweets by human users, there are a number of metaphor-related Twitter bots, which tweet automatically generated metaphors, or metaphor-like expressions, on a regular basis: see for example MetaphorIsMyBusiness (@MetaphorMagnet; @MetaphorMirror), run by the Creative Language System Group at University College Dublin (see Veale 2012). Metaphor-a-Minute (@MetaphorMinute) similarly tweets automatically generated ‘metaphors’, in the form of random ‘N is an N’ formulations. A typical example is ‘a gorilla is a shapelessness: bloody-minded yet categorized’ (16/10/16). Such Twitter feeds draw attention to the nature of metaphor and to the fine line it may straddle between profound wisdom and pure nonsense, in a manner not completely unlike some highly abstruse poetry.

More recently, online collections of metaphors like the Master Metaphor List have been put on a firmer footing with Andrew Goatly’s ‘Metalude’ project. This began in the 1990s

and has created an online database of conventional metaphors which can be searched by lexical item, root analogy (i.e. the underlying conceptual metaphor), source domain or target domain (www.ln.edu.hk/llc/cwd/project01/web/home.html). In addition to combing through lexicographical resources and existing research for examples of metaphor, the project gathered examples of metaphors submitted by users, in an example of crowdsourcing, *avant la lettre*. Procedures for submitting data are rigorous, with at least six lexical items, each of which meets a threshold number of corpus examples, being required to provide evidence of a proposed new root analogy.

Metalude is not the only online database of metaphor. A couple of quite different database resources have focused on metaphors of mental processes. Brad Pasenek's 'Mind is a Metaphor' site (<http://metaphors.iath.virginia.edu/>) incorporates a database of eighteenth-century metaphors of mind drawn from literary texts of all types and centred on the long eighteenth century in Britain. The database contains around 14,000 examples of metaphors identified by a variety of automatic and semi-automatic methods. These are categorised by a number of variables including period, genre and the gender, nationality, political affiliation and religion of the author. John Barnden's ATT-Meta Project Databank (www.cs.bham.ac.uk/~jab/ATT-Meta/Databank/) also concentrates on metaphors of the mind. This is an online databank of around 1100 examples of metaphorical descriptions of mental states and processes from texts, categorised according to the underlying conceptual metaphor. The ontology indicates when a metaphor is a special case of a more general metaphor (e.g. 'Ideas as possessions' is described as a quasi-special case of 'Ideas/emotions as physical objects') and includes a detailed description of the conceptual metaphor and links to examples from text.

The Amsterdam Metaphor Lab (<http://metaphorlab.org/>) is an example of a research centre with a well-established tradition of bringing metaphor and digital humanities methods together. The Metaphor Lab has created the VU Amsterdam Metaphor Corpus (www.vismet.org/metcor/search/), an online sample of around 190,000 words from the BNC-Baby corpus, annotated for metaphor using the MIPVU procedure (see also Steen et al. 2010) and including direct, indirect and implicit metaphor, as well as personification and signals that may point to metaphor (e.g. *as if* in the Amsterdam Metaphor Corpus example 'It is as if it is walking through a minefield', *so-called* and *like*). The Metaphor Lab has also compiled VisMet, a corpus of visual metaphors launched in 2014 and containing around 350 images, each annotated with source information, an interpretation of its meaning and information about any accompanying text (www.vismet.org/VisMet/). Opportunities for the study of visual and multimodal metaphor are also offered by the vast online archive of multimodal communication that is being collected by the Distributed Red Hen Lab – a co-operative of researchers who collaborate in the collection and sharing of data, and the development of theory, resources and tools (www.redhenlab.org).

Clearly there is a close relation between work in digital humanities and computational linguistics. Various tools exist that make it possible, to some extent, to identify metaphorical expressions automatically (e.g. Mason 2004; Berber Sardinha 2010; Neuman et al. 2013). While space precludes us from going into detail here, see Loenneker-Rodman and Narayanan (2012) for a recent overview of computational approaches to identifying metaphor in text and representing metaphor for use in natural language processing. One example of a prominent place for metaphor in commercial applications of computational linguistics was to be found in Yossarian (formerly at <https://yossarian.co/> and with further explanation on Wikipedia <https://en.wikipedia.org/wiki/YossarianLives>), a 'metaphorical search engine' that used artificial intelligence to return results that are potentially metaphorically related to the user's search term, with the aim of aiding creativity.

Metaphor in digital communication and environments

The large-scale, computer-aided study of metaphor, and other linguistic phenomena, has been facilitated by the fact that it is increasingly possible to download textual materials from the World Wide Web that previously only existed in hard copy, such as newspaper articles. This aspect of digital innovation, however, only affects the processes of data collection and analysis, rather than the use of metaphor itself in the data. In contrast, a growing number of studies focus on the use of metaphor in different types of digital environments. These studies show that the characteristics and possibilities of these environments can both facilitate new patterns in metaphor use and be shaped by metaphors that are selected for different design purposes.

Pihlaja (2014) investigates the role of metaphor in antagonistic YouTube videos on the topic of religion, known as YouTube ‘drama’. He shows how metaphors based on Bible stories are used, developed and challenged by YouTubers discussing Christian theology from different perspectives. For example, in late 2008 to early 2009 an argument developed between a Christian user called Yokeup and an atheist user called Crosisborg. Early in this argument, Yokeup insulted Crosisborg by describing him metaphorically as *human garbage*. This metaphor became the focus of heated discussion over a lengthy period of time and involved other users in addition to Yokeup and Crosisborg. The discussion centred primarily on whether the *human garbage* metaphor was legitimately based, as Yokeup claimed, on a particular parable from the New Testament, in which Jesus describes himself as a vine and human beings as branches that may remain attached to the vine or cut off and left to wither. In the course of the discussion, different users express their own positions through contrasting metaphorical narratives that develop in different ways the metaphorical idea of (some) people as *human garbage*. While the dynamic reuse, challenge and development of metaphors in interaction has been observed elsewhere (e.g. Cameron 2011), Pihlaja shows how the technical features and affordances of YouTube result in some phenomena that are typical of a particular kind of digital environment. For example, YouTube videos can persist for a long time but also be removed by the original user at any point. Yokeup did indeed remove the video in which he originally used the *human garbage* metaphor. As a result, some of the subsequent interactions and disagreements involved how his use of metaphor in the deleted video was remembered and reformulated by different users.

Pihlaja’s study also shows how metaphors can be used to insult and denigrate others. In this respect, online communication generally has provided more opportunities to observe offensive communicative behaviour, including by means of metaphorical uses of language. On the one hand, the geographical dislocation and relative anonymity of some types of online communication have a disinhibitory effect that leads some to engage in aggressive communication more than they may otherwise. On the other, these offensive messages persist online and can therefore become the subject of investigation. Demjén and Hardaker (2017), for example, show how offensive metaphors (e.g. *brainless scum*) were one of the strategies that were used on Twitter in 2013 to attack journalist Caroline Criado-Perez for successfully campaigning for the inclusion of the image of a female historical figure on UK banknotes.

Heino et al. (2010), in contrast, explore the metaphor of the ‘marketplace’ on an online dating site. The characteristics of such sites, they argue, are consistent with, and sometimes shaped by, this metaphor, as in the case of the ability to select particular characteristics to filter users’ profiles for closer consideration as potential romantic partners. Heino et al. analysed the use of market-related metaphors in 34 semi-structured interviews with users of a particular online dating site. Over half of the participants used economic metaphors for online dating without being prompted. For example, they described the site as a *supermarket*

and their own list of possible partners as a *sales pipeline*. What Heino et al. call the ‘market mentality’ is particularly evident in creative and extended uses of the ‘marketplace’ metaphor as in the following simile:

To me, [online dating is] like picking out the perfect parts for my machine where I can get exactly what I want and nothing I don’t want.

(Heino et al. 2010: 437)

Some interviewees, however, also questioned and sometimes resisted the metaphor, particularly because of how it may encourage people to make quick decisions about potential partners on the basis of superficial characteristics. In spite of these critical comments, Heino et al. (2010: 441) conclude that the design and functionality of online dating sites seem to encourage users to ‘adopt a marketplace orientation to the online dating experience’ and point out how this may result in the loss of a ‘holistic’ approach to other people in the process of looking for romantic partners.

Finally, Falconer (2008) provides an example of how metaphor can be used as a design concept in online learning environments. The challenge for such environments is how to organise content in such a way that users with different preferences and cognitive characteristics can be adequately catered for. Falconer describes how this challenge was met in one particular case by conceiving of content and resources on a learning site for students on a Distance Master’s course as a ‘universe’ consisting of different ‘constellations’. This metaphorical idea is then further developed:

To access the constellations in the RO [Research Observatory] universe the user enters a virtual building that is divided into seven ‘rooms.’ These rooms cover the basis of research, finding the focus, reviewing the literature, research methods, operationalizing research, analyzing the findings, and conclusions and reflections. Each of these rooms is visualized as containing a “telescope” that enables users to explore and interact with a range of constellations in a particular section of the universe that is relevant to the subject of the room.

(Falconer 2008: 121)

For example, the constellation ‘Defining aims’ for a student dissertation could be seen both from Room 1 (‘Finding the Focus’) and Room 3 (‘Reviewing the Literature’). The adoption of this cartographic metaphor in the design of the site was well-received by most students, as it made it easy to access the same content from different perspectives, and therefore catered successfully for different approaches to learning.

As these studies show, metaphor can be involved in different ways in digital communication and environments, not just as a way of speaking or writing but also as a conceptual and design tool.

Sample analysis

In this section we provide two concrete examples of digital humanities research on metaphor in English. The first example employs corpus methods and concerns the use of metaphor in an online forum for people with cancer. The second involves the creation and use of digital resources for the diachronic study of metaphor and concerns the conceptual domain REPTILES as both a source and target domain in the history of the English language.

Sample analysis 1 – Metaphors for cancer in an online patient forum

The ‘Metaphor in End-of-Life Care’ (MELC) project at Lancaster University (ESRC grant number: ES/J007927/1, 2012–14) involved the systematic analysis of metaphors in a 1.5-million-word corpus of interviews with and online forum posts by people with advanced cancer, family carers looking after a relative with cancer, and healthcare professionals. The corpus was analysed through a combination of qualitative manual analysis and quantitative corpus-based methods (Demmen et al. 2015; Semino et al. 2018). In particular, the approach included the use of the USAS semantic annotation tool in Wmatrix (Rayson 2008) to obtain semantic concordances, that is, lists of items classified under semantic tags that were likely to include a large proportion of metaphorical expressions (such as the semantic tags ‘Warfare’ and ‘Sports’). This resulted in the first large-scale systematic study of how the experience of cancer is talked about metaphorically in English by people based in the UK.

The findings of the analysis of the online patient data in particular have implications for ongoing debates about the appropriateness of military or, more generally, ‘violence’ metaphors, such as the *fight* against cancer. These metaphors have long been criticised as potentially harmful for patients, as they encourage an antagonistic relationship with the illness and may engender feelings of personal failure if the progression of the disease is described and perceived as the person *losing the fight*. The analysis of the MELC corpus did indeed provide evidence of the potentially harmful effects of violence metaphors generally, as when a person writing online about her incurable cancer says: *I feel such a failure that I am not winning this battle*. This is a clearly disempowering use of the metaphor, as it presents the sick person as helpless and ineffective and potentially responsible for her own impending death. Another contributor to the online forum makes this point very explicitly:

I sometimes worry about being so positive or feel I am being cocky when I say I will fight this as I think oh my god what if I don’t win people will think ah see I knew she couldn’t do it!

However, particularly when writing online, patients also use similar metaphors in empowering ways (i.e. to present themselves as active and determined) and to achieve a positive self-image, as in *Cancer and the fighting of it is something to be proud of* and *My consultants recognized that I was a born fighter* (Semino et al. 2017).

A similar pattern was observed for ‘journey’ metaphors for cancer, which have in some cases been suggested or adopted as preferable to violence metaphors (e.g. the UK’s 2007 NHS Cancer Reform Strategy: www.nhs.uk/NHSEngland/NSF/Documents/Cancer%20Reform%20Strategy.pdf; and the guidelines of the Cancer Research Institute of New South Wales: <https://www.cancer.nsw.gov.au/about-cancer-institute-nsw/media/writing-about-cancer-guidelines>). When writing online, some patients use journey metaphors in empowering ways (e.g. *I passionately believe we all need to choose our own road through this*), while others use them to present themselves as disempowered (e.g. *How the hell am I supposed to know how to navigate this road I do not even want to be on when I’ve never done it before*). Overall, this study therefore shows that, when dealing with cancer, there are no metaphors that can be straightforwardly banned as harmful or promoted as helpful for everyone. Rather, each type of metaphor can be used in different ways by different people in different contexts, particularly in terms of emotional associations and (dis)empowerment. Therefore, the best way to support patients is to enable and encourage them to use the metaphors that work best for them. This means that each person should be in a position to draw from different metaphors at different points in their

experience of illness. To this end, the findings of the study have led to the development of a ‘Metaphor Menu’ for people with cancer – a list of quotes including a wide variety of different metaphors for the experience of cancer, which can ideally be used to provide resources for communication and sense-making for people with new diagnoses or at different points in their experience of the illness (see Semino 2014; <http://wp.lancs.ac.uk/melc/the-metaphor-menu/>).

Mutual support through metaphor on the online patient forum

Several studies have pointed out that peer-to-peer online forums devoted to illness can provide a source of support and sense of community for contributors (e.g. Allen et al. 2016). The analysis of the online patient section of the MELC corpus shows that emotional support can be offered by means of metaphors.

In some cases, contributors reuse one another’s metaphors in ways that suggest that they have similar views and feelings. For example, one forum contributor metaphorically describes his approach to the first round of chemotherapy in terms of a race (NB: The relevant expressions are underlined; original spellings are retained):

I start my first round of chemo tomorrow . . . and just want to get things started. I am keeping focused on the tape at the end of this race so whatever happens on the way is only a way to get there.

Another contributor then responds with an encouraging post that develops the metaphor further:

It’s good you have this first hurdle out of the way, . . . You have a very long road ahead so pace yourself and try to be patient, treat it like a marathon but in the end you will get there.

This partial reuse and extension of another’s metaphor has been described as one of the ways in which empathy is conveyed in interaction (Cameron 2011), and can in fact be seen as the opposite of what Pihlaja (2014) noticed in the YouTube drama on the *human garbage* metaphor.

In other cases, contributors may introduce a particular metaphor as a way of helping another person view their own situation in a more positive way. For example, the following extract is part of a response to someone who had just been diagnosed with a form of cancer that was likely to be incurable:

Sorry to learn of your diagnosis, . . . if you can think of this chapter in your life as a reluctant journey and each procedure a place along that journey that must be completed before you move onto the next it may help you better deal with it.

Both the metaphorical use of *chapter* and the extended journey metaphor are used to suggest that even a potentially terminal diagnosis should be seen as part of a longer process with many stages, so that it is possible and preferable for the person to focus on one part of the process at a time. This metaphorical framing is explicitly offered to the addressee as a potentially beneficial way of conceiving of the situation (*if you can think . . . it may help you better deal with it*).

Overall, the computer-aided analysis of the online patient corpus has led to qualitative observations that confirm earlier findings on online interactions about illness and reveal the

role that metaphors can have in such interactions, particularly in terms of the expression of emotions, self-perceptions and interpersonal relationships.

Sample analysis 2 – Reptile metaphors in the ‘Metaphor Map’

The second study that we illustrate here is the ‘Mapping Metaphor with the Historical Thesaurus’ project, which was carried out at the University of Glasgow between 2012 and 2016 (AHRC grant number: AH/I02266X/1; <http://mappingmetaphor.arts.gla.ac.uk/>). This digital humanities project used semi-automated methods to chart the presence of metaphor across the entire history of the English language (from the Old English period to the present day), taking up the challenge from Kövecses that ‘[m]ore precise and more reliable ways of finding the most common source and target domains are needed’ (2010: 27). The focus is on metaphor in the language system as recorded in the major lexicographical resource of the Historical Thesaurus of English, also compiled at Glasgow, over a period of 40 years. It contains nearly 800,000 word forms, organised by meaning into a hierarchical system of nearly a quarter of a million categories and subcategories. The Historical Thesaurus itself combines the evidence of the *Oxford English Dictionary* (second edition) with that of *A Thesaurus of Old English* (Roberts and Kay 1995) for words restricted to the earlier period c700–1150AD. While the particular focus of Mapping Metaphor is on metaphor as recorded in the language system, that record draws on evidence of metaphor in text (the OED citations) and ultimately – it can be presumed – from metaphor in the minds of speakers of English.

At the root of the project methodology is the fact it is possible to see metaphor in the polysemy that manifests itself in a semantically organised set of data like a thesaurus. For example, a polysemous word like *butterfly* appears in several places in a thesaurus categorisation system. In the Historical Thesaurus, disregarding part of speech, it appears in 13 categories and subcategories, including of course under ‘Invertebrates’, under ‘Constitution of matter’ for ‘a weak substance or thing’, under ‘Inattention’ to denote a light-minded person and under ‘Social event’ with the verbal sense ‘participate in social events’. In this case, the last three are clearly metaphorical extensions of the first, highlighting the physical or perceived mental flimsiness of the invertebrate or its characteristic fast fluttering movement.

While computers can very readily identify the lexis that semantic categories share, they are not very effective at distinguishing metaphor from other types of polysemy, or from homonymy, so the project’s initial automatic analysis was followed by a lengthy stage of manual analysis (see also Alexander and Bramwell 2014 for an overview of the project methodology). The result of this analysis is the ‘Metaphor Map of English’ and ‘Metaphor Map of Old English’ online resources, visualisations which show all of the metaphorical connections identified between pairs of semantic categories over the history of English (see Figure 8.1). The Metaphor Maps also indicate the source and target of each metaphor, give examples of words instantiating the metaphor and are coded as ‘Strong’ or ‘Weak’ to indicate the coder’s judgement of the systematicity of the connection, based on the lexical evidence from the Historical Thesaurus.

A significant benefit of the Metaphor Map methodology is that it allows for systematic analysis of both source and target domains. It is relatively straightforward to establish the source of metaphors within a chosen target domain: these will emerge from a glance through the relevant section of a good thesaurus. Thus, on even a quick skim through the category ‘Lack of understanding’ in the Historical Thesaurus, the extent of speakers’ reliance on ideas

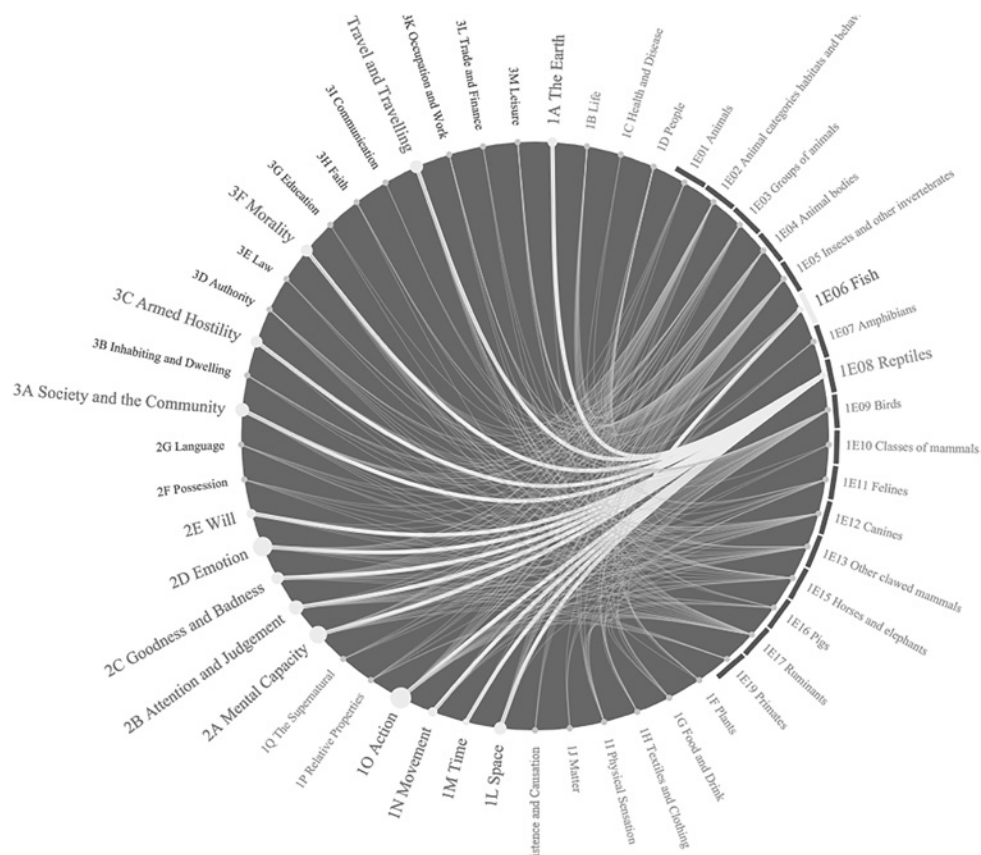


Figure 8.1 Metaphor Map of English showing connections between category 1E08 Reptiles and all of the macro-categories

of density and slowness to convey concepts in this semantic field becomes clear. But this is not so simple when we are interested in the semantic domains in which such concepts act as the source. Here, semi-automated analysis tracking lexis across all semantic domains is required. We illustrate this through a sample analysis of category 1E08 Reptiles in the Metaphor Map data. This category falls within Part 1, the External World in the Metaphor Map, and within the macro-category 1E Animals. For as full as possible a picture of the extent and role of metaphor relating to the Reptiles category, we begin with an overview of concepts from this category as the target of metaphor and then consider the category as source.

Reptiles as target

There is no evidence from Old English of Reptiles as target. Evidence starts to appear from the early fifteenth century onwards, and a total of nine category pairs instantiate a metaphorical connection in which Reptiles acts as target. These are shown in Table 8.1, which indicates each source category and gives a lexical example to illustrate the metaphorical connection. In seven of the nine, in fact, Reptiles is both target and source; in particular, a number of categories in the broad areas of warfare are bidirectional with regard to metaphor.

Table 8.1 Metaphorical connections with 1E08 Reptiles as target

Source category	Illustrative words (with date of transferred sense from OED2 and gloss of sense in target category)
1E11 Felines	<i>tiger</i> (1895, species of snake with banded markings)
1F05 Cultivated plants	<i>millet</i> (1608–1661, species of snake with protuberant markings like millet seed)
1O15 Safety	<i>monitor</i> (1826 –, type of lizard); <i>safeguard</i> (1831–1847/9, type of lizard, said to give warning of the approach of alligators)
2B12 Beauty and ugliness	<i>comb</i> (c1400–1661, serrated crest on some serpents)
2F10 Taking and thieving	<i>fang</i> (1800 –, venom tooth of serpent, transferred from <i>fang</i> , an instrument for catching or holding)
3C01 War and armed hostility (same connection also present in 3C02 Military forces; 3C03 Weapons and armour)	<i>spear</i> (1608 – a1700, the sting of a reptile or insect); <i>soldier</i> (1608, sea-tortoise of the order Chelonia named for protective shell)
3J02 Transport	<i>coach-whip</i> (1827 –, snake of the genus Masticophis, which resembles a black-handled coach-whip)

These metaphorical connections can all be considered forms of ‘resemblance metaphor’, a type proposed by Grady (1999), which includes image metaphors and stands in contrast to ‘correlation metaphors’ (that is, metaphors of the classic LIFE IS A JOURNEY type). More recently, Ureña and Faber (2010) have refined this model through a study of metaphors in the domain of marine biology, showing that the category of resemblance metaphors is graded, encompassing image metaphors and behaviour-based metaphors, with the degree of dynamicity of the underlying mental image accounting for different types of metaphor within the category.

Ureña and Faber’s findings are echoed in the data in Table 8.1, which shows those metaphors in which an aspect of visual appearance or behaviour is transferred to a concept in the category of Reptiles. The aspects of appearance include markings and the shape and hardness of body parts. *Safeguard* illustrates a transfer of behavioural qualities (as presumably does *monitor*, although the motivation for its meaning is less clear). We can also see the graded nature of resemblance metaphors in examples like *coach-whip*, which combines shape (appearance) with movement (behaviour).

Reptiles as source

Not surprisingly, given the strong tendency for metaphors to have a source that is more concrete than the target, concepts from within the External World category of Reptiles can be seen much more commonly to be the source in metaphorical connections. Here, therefore, we will have to be selective in our treatment of the scope of such metaphors.

The Metaphor Map of Old English shows category 1E08 Reptiles as the source for metaphorical connections with six other categories (see also Hough 2017, for a consideration of these in the context of other animals in this period, and Dallachy 2016 for some comments on *snake* in the context of thievery). Four of these are united by the lexical item *wyrm* (or its derived form *wyrmlic*), meaning a serpent, snake or dragon in Old English, though also with the more general sense of ‘any animal that creeps or crawls’ including insects, from which the present-day literal meaning stems. This is transferred to 1O18 Adversity (‘wretched person’), 1Q06 Hell (‘torment of hell’, from the sense ‘huge, like a dragon’), 2B08 Contempt (‘contemptible person’) and

2D06 Emotional suffering ('grief that preys on the heart'). The link between Reptiles and 1Q04 Devil is instantiated by *draca* ('serpent, dragon', representing the Devil).

These connections cluster within the conceptual fields of the supernatural and emotional states, broadly conceived. This tendency continues in the later period, with many more metaphorical connections emerging; the Metaphor Map of English shows 61 categories as the target with a source in 1E08 Reptiles. For the supernatural, category 1Q04 Devil sees the arrival of *serpent* and *serpentine* in the early fourteenth century. The connection with 2B08 Contempt is strengthened by the addition of lexical items such as *reptile* and *hiss*, and that with 2D06 Emotional suffering is similarly strengthened by *atter* and *side-winder*. Related concepts in categories such as 2D07 Anger and 2D12 Jealousy also draw on Reptiles.

We saw earlier examples of Reptiles as the target of metaphorical connections with categories in the semantic domain of war and hostility (3C01 War and armed hostility, 3C02 Military forces and 3C03 Weapons and armour). Some interesting bidirectional connections appear here, with Reptiles also contributing as the source, through lexical items such as *tortoise-shell*, *tortoise*, *testudo* (a shield-wall), *snake* (sergeant in Australian military slang) and *serpentine* and *basilisk* (large or long pieces of artillery).

To give an indication of the extent of such metaphors, Table 8.2 shows some examples of metaphorical connections instantiated by words for concepts other than generic snakes and tortoises. Finally, as we would expect, extremely well-attested connections emerge between Reptiles and categories such as 1O22 Behaviour and conduct, 2C02 Bad, 3F01 Morality and immorality and 3F05 Moral evil, with a raft of *vipers*, *reptiles*, *serpents* and *snakes* showing a long-standing metaphorical connections between reptiles and different types of badness.

Category 1E08 Reptiles is only one of 415 semantic categories in the Metaphor Map resource; indeed, it is one of the smaller categories. Nevertheless, it can begin to show some of the possibilities of this digital humanities resource for deepening our understanding of the English language in its historical and present-day forms. Further case studies of semantic domains drawing on the Metaphor Map data can be found in Anderson et al. (2016).

Critical issues and future directions

It is undeniable that the digital turn in research has impacted very significantly on the study of the English language, as this volume as a whole makes clear. Within this, as this chapter

Table 8.2 Sample metaphorical connections with 1E08 Reptiles as source

Target category	Illustrative words (with date of transferred sense from OED2 and gloss of sense in target category)
1A05 Landscape, high and low land	<i>turtle-back</i> (1913 –, rising ground)
1J34 Colour	<i>chameleon</i> (1687, changing colour)
1L03 Size and spatial extent	<i>pythonic</i> (1860 –, huge); <i>tyrannosaur(us)</i> (1957 –, that which is huge); <i>dinosaur</i> (1979 –, that which is huge)
1M07 Measurement of time and relative time	<i>brontosaurian</i> (1909–1977, old-fashioned); <i>dinosaur</i> (1952 –, that which is old-fashioned)
2E05 Decision-making	<i>chameleon</i> (1586 –, inconstant person)
3F07 Licentiousness	<i>cockatrice</i> (1599–1747, unchaste woman); <i>lizard</i> (1935, licentious male)
3J01 Travel and journeys	<i>crocodile</i> (1870 –, school pupils walking in order)

has shown, research on metaphor that sits under the ‘big tent’ of digital humanities (Terras 2013: 263) is diverse. Computational linguistics and natural language processing can be found alongside corpus linguistics and novel approaches to historical semantics. Such work offers insights into metaphor in text (whether ‘traditional’ or computer-mediated), in the language system and in the conceptual system, as well as a perspective on the very limits of metaphor (for example, Metaphor-a-Minute).

Perhaps stemming from the traditional focus that the precursors to digital humanities placed on the study of written language, there is still a text bias in work on metaphor. There is a strong concentration of work on verbal metaphor, though notable exceptions do exist, for example in the pioneering work of the Amsterdam Metaphor Lab, including the VisMet corpus of visual metaphor, which combines image and text. As digital techniques for handling visual images are refined, we can expect to see more systematic work on visual metaphor in the near future, including also perhaps on metaphor as it manifests itself in gesture.

Similarly, there is certainly more to be learned about metaphor as it both reflects and shapes the characteristics and affordances of digital communication. The Metaphor in End-of-Life Care project and Pihlaja’s work on YouTube drama offer insights into the rich possibilities here. But as social media plays an increasingly significant part in how we form our views on everything from fashion to politics, there is much still to explore with regard to the place of metaphor in conveying elusive meaning.

Finally, we have as yet made relatively little progress on automatic metaphor identification. While concordance software enables easy identification of pre-selected lexical items from the source or target domain, or of metaphor signalling expressions (*as if*, *like* and *literally*, for example), mechanisms for *systematically* identifying all of the metaphor in a text or corpus without prior knowledge of its particular surface manifestations are still far off, despite the opportunities brought about by the USAS semantic annotation tool in Wmatrix (Rayson 2008), as used in the ‘Metaphor in End-of-Life Care’ project. This is surely a key direction for future research, and the fuller understanding of metaphor that projects like ‘Metaphor in End-of-Life Care’ and ‘Mapping Metaphor with the Historical Thesaurus’ give us will only speed progress.

Some digital humanities scholars, such as Robinson (2014: 255), look ahead to a time, likely not far off, when ‘the prefix “digital” . . . will be redundant, once so much of the humanities is digital’. Metaphor scholars have been particularly enthusiastic in their adoption of digital methods, and the digital environment has both supported and provided direction for research on metaphor in recent decades. Some groundwork has been laid. However, given that scholarly interest in metaphor has reached a level of maturity at broadly the same time as we have seen increasing sophistication of digital – and digital humanities – methods, it is perhaps surprising that there has not been an even greater exploitation of digital methods for the study of metaphor. Certainly, there remain many questions to be answered and many opportunities for researchers. The road that we find ourselves on, following the digital turn, will no doubt be a long one.

Further reading

- 1 Anderson, W., Bramwell, E. and Hough, C. (eds.) (2016). *Mapping English metaphor through time*. Oxford: Oxford University Press.

This is a volume of essays stemming from the ‘Mapping Metaphor with the Historical Thesaurus’ project. Each chapter presents a data-driven case study of a semantic domain or small set of related semantic domains, such as LANDSCAPE, EXCITEMENT, COLOUR and FEAR.

- 2 Deignan, A. (2005). *Metaphor and corpus linguistics*. Amsterdam: John Benjamins.
Through lots of detailed examples, Deignan demonstrates the advantages of corpus data for investigating metaphor, evaluating this method alongside other complementary approaches, such as cognitive and discourse approaches to metaphor.
- 3 Pihlaja, S. (2014). *Antagonism on YouTube: Metaphor in online discourse*. London: Bloomsbury.
This book concentrates on the use of metaphor in the construction of online arguments on religious topics. Pihlaja carries out a detailed discourse analysis of YouTube videos and user interaction.
- 4 Semino, E., Demjén, Z., Demmen, J., Koller, V., Payne, S., Hardie, A. and Rayson, P. (2017). The online use of Violence and Journey metaphors by patients with cancer, as compared with health professionals: A mixed methods study. *BMJ Supportive & Palliative Care* 7: 60–66.
This article presents the methodology and findings of a major part of the ‘Metaphor in End-of-Life Care’ project (see earlier) for an audience of healthcare professionals. It shows how corpus linguistic methods were used to study the use of violence and journey metaphors on an online forum for people with cancer and to compare their frequencies with those found on an online forum for healthcare professionals.

References

- Alexander, M. and Bramwell, E. (2014). Mapping metaphors of wealth and want: A digital approach. *Studies in the Digital Humanities* 1. Available at: www.hrionline.ac.uk/openbook/chapter/dhc2012-alexander (Accessed February 2018).
- Allen, C., Vassilev, I., Kennedy, A. and Rogers, A. (2016). Long-term condition self-management support in online communities: A meta-synthesis of qualitative papers. *Journal of Medical Internet Research* 18(3): e61.
- Anderson, W. (2013). ‘Snippets of memory’: Metaphor in the SCOTS corpus. In W. Anderson (ed.), *Language in Scotland: Corpus-based studies*. Amsterdam and New York: Rodopi, pp. 215–236.
- Anderson, W., Bramwell, E. and Hough, C. (eds.) (2016). *Mapping English metaphor through time*. Oxford: Oxford University Press.
- Berber Sardinha, T. (2010). A program for finding metaphor candidates in corpora. *The Especialist (PUCSP)* 31: 49–68.
- Cameron, L. (2011). *Metaphor and reconciliation: The discourse dynamics of empathy in post-conflict conversations*. London: Routledge.
- Cameron, L. and Deignan, A. (2003). Combining large and small corpora to investigate tuning devices around metaphor in spoken discourse. *Metaphor and Symbol* 18(3): 149–160.
- Charteris-Black, J. (2001). Blood, sweat and tears: A corpus-based cognitive analysis of ‘blood’ in English phraseology. *Studi Italiani di Linguistica Teorica e Applicata* 30(2): 273–287.
- Charteris-Black, J. (2004). *Corpus approaches to critical metaphor analysis*. Basingstoke: Palgrave Macmillan.
- Dallachy, F. (2016). The dehumanized thief. In W. Anderson, E. Bramwell and C. Hough (eds.), *Mapping English metaphor through time*. Oxford: Oxford University Press, pp. 208–220.
- Deignan, A. (2005). *Metaphor and corpus linguistics*. Amsterdam: John Benjamins.
- Deignan, A., Littlemore, J. and Semino, E. (eds.) (2013). *Figurative language, genre and register*. Cambridge: Cambridge University Press.
- Deignan, A. and Potter, L. (2004). A corpus study of metaphors and metonyms in English and Italian. *Journal of Pragmatics* 36(7): 1231–1252.
- Demjén, Z. and Hardaker, C. (2017). Metaphor, impoliteness and offence in online communication. In E. Semino and Z. Demjén (eds.), *The Routledge handbook of metaphor and language*. London and New York: Routledge, pp. 353–367.
- Demmen, J., Semino, E., Demjén, Z., Koller, V., Hardie, A., Rayson, P. and Payne, S. (2015). A computer-assisted study of the use of Violence metaphors for cancer and end of life by patients, family carers and health professionals. *International Journal of Corpus Linguistics* 20(2): 205–231.

- Falconer, L. (2008). Evaluating the use of metaphor in online learning environments. *Interactive Learning Environments* 16(2): 117–129.
- Fauconnier, G. and Turner, M. (2002). *The way we think: Conceptual blending and the mind's hidden complexities*. New York: Basic Books.
- Grady, J.E. (1999). A typology of motivation for conceptual metaphor: Correlation vs resemblance. In R.W. Gibbs and G.J. Steen (eds.), *Metaphor in cognitive linguistics*. Amsterdam and Philadelphia: John Benjamins, pp. 79–100.
- Heino, R.D., Ellison, N.B. and Gibbs, J.L. (2010). Relationshipshopping: Investigating the market metaphor in online dating. *Journal of Social and Personal Relationships* 27(4): 427–447.
- Hough, C. (2017). A new perspective on Old English metaphor. In C. Biggam, C. Hough and D. Izdebska (eds.), *The daily lives of the Anglo-Saxons*. Essays in Anglo-Saxon Studies 8. Tempe: Arizona Center for Medieval and Renaissance Studies.
- Johansson Falck, M. and Gibbs, R.W. Jr. (2012). Embodied motivations for metaphorical meanings. *Cognitive Linguistics* 23(2): 251–272.
- Kirschenbaum, M.G. (2013). What is digital humanities and what's it doing in English departments? In M. Terras, J. Nyhan and E. Vanhoutte (eds.), *Defining digital humanities: A reader*. Farnham: Ashgate, pp. 195–204.
- Koller, V. (2004). *Metaphor and gender in business media discourse: A critical cognitive study*. Basingstoke: Palgrave Macmillan.
- Kövecses, Z. (2000). *Metaphor and emotion: Language, culture, and body in human feeling*. Cambridge: Cambridge University Press.
- Kövecses, Z. (2010). *Metaphor: A practical introduction* (2nd ed.). Oxford: Oxford University Press.
- Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (ed.), *Metaphor and thought*. Cambridge and New York: Cambridge University Press, pp. 202–251.
- Lakoff, G., Espenson, J. and Schwartz, A. (1991). *Master metaphor list*. Second draft copy. Technical report, Cognitive Linguistics Group, University of California Berkeley. Available at: <http://cogsci.berkeley.edu>; now accessible from <http://araw.mede.uic.edu/~alansz/metaphor/METAPHORLIST.pdf> (Accessed February 2018).
- Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
- Lakoff, G. and Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York: Basic Books.
- L'Hôte, E. (2014). *Identity, narrative and metaphor: A corpus-based cognitive analysis of New Labour discourse*. Basingstoke: Palgrave Macmillan.
- Loenneker-Rodman, B. and Narayanan, S. (2012). *Computational approaches to figurative language*. In M. Spivey, K. McRae and M. Joannisse (eds.), *The Cambridge handbook of psycholinguistics*. Cambridge: Cambridge University Press, pp. 485–504.
- Mason, Z.J. (2004). CorMet: A computational, corpus-based conventional metaphor extraction system. *Computational Linguistics* 30(1): 23–44.
- McCarty, W. (2013). Tree, turf, centre, archipelago – or wild acre? Metaphors and stories for Humanities Computing. In M. Terras, J. Nyhan and E. Vanhoutte (eds.), *Defining digital humanities: A reader*. Farnham: Ashgate, pp. 97–118.
- McEnery, T. and Hardie, A. (2012). *Corpus linguistics: Methods, theory and practice*. Cambridge: Cambridge University Press.
- Musolff, A. (2004). *Metaphor and political discourse: Analogical reasoning in debates about Europe*. Basingstoke: Palgrave Macmillan.
- Neuman, Y., Assaf, D., Cohen, Y., Last, M., Argamon, S., Howard, N. and Frieder, O. (2013). Metaphor identification in large texts corpora. *PloS One* 8(4): e62343.
- Pihlaja, S. (2014). *Antagonism on YouTube: Metaphor in online discourse*. London: Bloomsbury.
- Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics* 13(4): 519–549.

- Rieder, B. and Röhle, T. (2012). Digital methods: Five challenges. In D.M. Berry (ed.), *Understanding the digital humanities*. Houndmills: Palgrave Macmillan, pp. 67–84.
- Roberts, J., Kay, C. and Grundy, L. (1995). *A thesaurus of old English*. (= King's College London Medieval Studies XI.) (2nd ed.). Amsterdam: Rodopi.
- Robinson, P. (2014). Digital humanities: Is bigger, better? In P.L. Arthur and K. Bode (eds.), *Advancing digital humanities*. Basingstoke: Palgrave Macmillan, pp. 243–257.
- Salahshour, N. (2016). Liquid metaphors as positive evaluations: A corpus-assisted discourse analysis of the representation of migrants in a daily New Zealand newspaper. *Discourse, Context and Media* 13: 73–81.
- Semino, E. (2014). A metaphor menu for cancer patients'. *eHospice*. Available at: www.ehospice.com/uk/articleview/tabid/10697/articleid/13414/language/en-gb/view.aspx.
- Semino, E., Demjén, Z., Demmen, J., Koller, V., Payne, S., Hardie, A. and Rayson, P. (2017). The online use of Violence and Journey metaphors by patients with cancer, as compared with health professionals: A mixed methods study. *BMJ Supportive & Palliative Care* 7: 60–66.
- Semino, E., Demjén, Z., Hardie, H., Payne, S. and Rayson, P. (2018). *Metaphor, cancer and the end of life: A corpus-based study*. New York: Routledge.
- Simó, J. (2011). Metaphors of blood in American English and Hungarian: A cross-linguistic corpus investigation. *Journal of Pragmatics* 43(12): 2897–2910.
- Skorczynska, H. (2010). A corpus-based evaluation of metaphors in a business English textbook. *English for Specific Purposes* 29(1): 30–42.
- Skorczynska, H. and Deignan, A. (2006). A comparison of metaphor vehicles and functions in scientific and popular business corpora. *Metaphor and Symbol* 21(2): 87–104.
- Steen, G.J., Dorst, A.G., Herrmann, J.B., Kaal, A.A., Krennmayr, T. and Pasma, T. (2010). *A method for linguistic metaphor identification. From MIP to MIPVU*. Amsterdam: John Benjamins.
- Stefanowitsch, A. (2006). Words and their metaphors: A corpus-based approach. In A. Stefanowitsch and S. Th. Gries (eds.), *Corpus-based approaches to metaphor and metonymy*. Berlin: Mouton de Gruyter, pp. 63–105.
- Terras, M. (2013). Peering inside the big tent. In M. Terras, J. Nyhan and E. Vanhoutte (eds.), *Defining digital humanities: A reader*. Farnham: Ashgate, pp. 263–270.
- Terras, M. (2016). A decade in digital humanities. *Journal of Siberian Federal University Humanities & Social Sciences* 7: 1637–1650.
- Ureña, J.M. and Faber, P. (2010). Reviewing imagery in resemblance and non-resemblance metaphors. *Cognitive Linguistics* 21(1): 123–149.
- Veale, T. (2012). *Exploding the creativity myth: The computational foundations of linguistic creativity*. London: Bloomsbury.

Grammar

Anne O’Keeffe and Geraldine Mark

Introduction

Grammar, according to Carter and McCarthy, in its simplest description, consists of two basic principles: ‘the arrangement of items (syntax) and the structure of items (morphology)’ (2006: 2). However, as Swan notes, ‘simple in principle, grammar generates considerable complexity in practice’ (2011: 558). It is widely recognised that restricting a definition of grammar to its basic principles of syntax and morphology tells us only about its form, or structure, but not its usage. Grammar is never context-free and does not exist in a vacuum. As Larsen-Freeman notes ‘grammar *is* about the form of the language, but it is also used to make meaning’ (2014: 256). Its forms have meanings, and these meanings are put to use.

Carter and McCarthy (2006) point to the difference between grammar as structure and grammar as choice. They note that when we speak and write, we make choices from a range of options available to us to create meaning and that the grammatical choices we make contribute to these meanings. These choices are dependent not just on the surrounding words, phrases, clauses, sentences and paragraphs but on whole texts, or strings of discourse. Additionally, these choices reflect our surroundings, the time and place where the text or discourse is created and, crucially, the relationships between participants. We use grammar to create relationships, to interact and convey messages and meanings for our own communicative purposes.

Grammar, therefore, is at the disposal of its users. It is an ever-changing and dynamic resource. Many definitions of grammar include the words ‘rules’ ‘correct use’ and ‘accepted principles’ and are framed with references to prescription and proscription. Grammar can be a highly contentious issue, and it is bound up with issues of acceptability and correctness. As Carter and McCarthy (2017: 2) aptly put it, ‘grammar has long been a source of controversy, with age-old themes rearing their heads periodically’. What is important to note is that some rules and principles are stable while others are dynamic and subject to change and, as this chapter will detail, the advent of digital tools within the field of corpus linguistics has fundamentally changed how we examine and describe grammar. We will begin with a historical overview which traces how the study of grammar moved from a manual process of analysing, investigating and describing language to a more technologically aided process.

To begin, grammar books were the result of painstaking manual examination and introspection. While early printing technology brought a proliferation of prescriptive grammar books, the models and paradigms were inherited from existing works. Eventually, with the advent of audio recording, text scanning and large-scale corpora, more recent digital technology changed the way we research and describe grammar.

Beginning our chapter in a pre-technology era may seem out of place in a volume on digital humanities (DH) but it is important to understand the lineage (see Terras 2016). DH and English departments have always been a likely partnership, given that texts can be very easily digitised and manipulated by computers (Kirschenbaum 2010), and so from an early stage in the development of computing we can track a lineage of work on texts using computers in areas that include stylistics, corpus linguistics and authorship studies. Terras (2016) goes as far as saying that linguistics is a prime example of the broader notion of the DH. (This is not surprising given that being able to investigate and analyse language on an ever-growing scale can only be of benefit to the study of language.) And yet, as Terras (*ibid*) points out, the investigation and analysis of texts is not new. While technological advances in computing allow us to manipulate texts so as to count and concordance words in any given collection or corpus of texts, counting and concordancing were a common scholarly practice going back to the 1500s when alphabetised concordances and frequency tables were made on occurrences of words in the Bible (McCarthy and O’Keeffe 2010).

Quantification was an established method in humanities, as Terras (2016) reflects, and while DH was only a relatively recent coinage, using computers in the humanities across a range of disciplines has been there for decades. The umbrella term DH seems to be a useful moniker that can be retrofitted to this type of endeavour. Yet, Terras (*ibid*), among others, would argue that it has more significance than being a useful catch-all. She argues that by looking at the use of quantification and manipulation of texts in a pre- and post-computer era in the study of humanities, we should see DH as part of a lineage and part of a natural progression in quantitative methods that have evolved over the past 500 years. We will now take a look at the emergence of the study of English grammar in this longitudinal context so as to place DH as part of its developmental process. We note that DH is not an endpoint in this development but that concerted developments in DH, especially in relation to computing, can ultimately advance our ability to investigate grammar.

Background

Historical development of grammars of English

There has been a long tradition of grammar writing over the centuries in English. We can trace the lineage of grammars of English back to the 1500s, and because of its association with godliness and learnedness, Latin had a great influence on early grammar books. Linn (2006) notes that the Latin grammar cast a long shadow on the early history of English grammars and there was a desire to match the grammars of Latin which were used at the time (Linn 2006; Carter and McCarthy 2017). In fact, many early grammars of English were written in Latin, as this was seen as more scholarly (Linn 2006). Often, they were also written by clergymen.

As noted by Linn (2006) and McCarthy (2017a), Bullokar’s *Pamphlet for Grammar*, published in 1586, is widely believed to be the first proper grammar of English. In it, we see that the eight basic parts of speech that we are familiar with today were well established 400 years ago (McCarthy 2017a). This, and grammars which followed down the centuries,

described English grammar within a Latin framework which was part of an effort to show that English, like Latin, was rule-bound (Linn 2006) even though this was not always the case in reality. As both Linn (2006) and Aitchison (2001) point out, any deviation from the description of a rule which did not fit with the Latin grammar paradigms was seen as an untruth (and possibly sinful).

Also of note in relation to early grammars is that they were written for the privileged, so-called learned, male reader (Linn 2006). Interestingly, with growth in British trade and commerce towards the second half of the seventeenth century, there was an increase in the demand for grammars to be written for an audience outside of Britain, especially those learning English (Linn 2006). This is reflected in the fact that by the end of the seventeenth century, there were 270 new grammars (Linn 2006). The most widely-known eighteenth-century grammar was *A Short Introduction to English Grammar, with critical notes* (1762), written by clergyman Robert Lowth. Howatt and Widdowson (2004) state that though Lowth claims to have written the grammar for ‘private and domestic use’ (quoted in Howatt and Widdowson 2004: 116), it went on to ‘make him the most famous and emulated grammarian of his time’ (ibid: 116), and his influence extended into the nineteenth century through his disciples Murray and Cobbett (ibid).

Howatt and Widdowson (2004) document that, notably, Lowth went to great lengths to include many examples as well as common errors and that this contributed to the book’s widespread use and usefulness. This made it especially suited to independent study, unlike many other grammars of the time, and many see it as the first pedagogic grammar (Howatt and Widdowson 2004). Though they say there is no evidence to support it, Howatt and Widdowson (2004: 120) point out that the main criticism of Lowth’s focus on errors was that he invented them and that they derived solely from ‘his own assumptions and prejudices’. At the time, Lowth caused much controversy because he was not shy of pointing out errors in prestigious literary works at the time (e.g. Pope and Swift). Aitchison (2001: 11) refers to Lowth’s 1762 grammar as having ‘a surprising influence, perhaps because of his own high status’ and asserts that many grammars in use today have ‘laws of “good usage” which can be traced directly to Bishop Lowth’s idiosyncratic pronouncements as to what was ‘right’ and what was ‘wrong’.

McCarthy (2017a) lists the work of Murray, one of Lowth’s disciples, entitled *English Grammar, Adapted to the different classes of learners* (1795, 1st Ed., 1809 16th Ed.) as one of the most popular and influential works. The interesting shift in Murray’s work is that it is targeted at the masses. This trend continues in line with massification in education at the time, leading to some fascinating titles such as Edward Shelley’s *The People’s Grammar; or English Grammar without Difficulties for ‘the million’* (1848). Cobbett’s work, *A Grammar of the English Language, In a Series of Letters* (originally published in 1818) had the intriguing sub-title, *Intended for the Use of Schools and of Young Persons in General; But More Especially for the Use of Soldiers, Sailors, Apprentices, and Plough-boys. To which are Added Six Lessons Intended to Prevent Statesmen from Using False Grammar and from Writing in an Awkward Manner*.

These grammars carried influence within the British and American schooling systems right into the twentieth century (Linn 2006), and what is central is that they had a moral tone as to the propriety of good grammar. Counter to this, there was a societal intolerance towards ‘bad grammar’. They prescribed grammar rules as absolute truths and proscribed what ought not to be said or written.

Linn (2006) points to an important shift in how grammar was described in the nineteenth century to a paradigm which still pertains today. This was their movement away from

solely prescribing grammar within the traditional (Latin-based) ‘word-and-paradigm’ model which focused solely how words related to other words (Linn 2006). Influenced by the work of German grammarian, Karl Ferdinand Becker, there was a shift towards looking at how words related to grammatical units, within a clause-based paradigm.

The next major turn in how grammar was described came in towards the end of the nineteenth century and was very much enhanced by the introduction of audio-recording technology. Through close observation of the spoken word some attention was now given to spoken grammar (Linn 2006; Carter and McCarthy 2017). This shift is attributed to developments within the field of phonetics, most importantly, the work of Sweet and his grammars *The Practical Study of Languages: a guide for teachers and learners* (1899) and *A new English Grammar: logical and historical. Part I: introduction, phonology, and accidence* (1900). What these works achieved was the importance of spoken language as an integral part of describing the grammar of a language (Linn 2006; Carter and McCarthy 2017). Smith (1990: 80) refers to Palmer’s (1924) *A Grammar of Spoken English* as a ‘comprehensive and ground-breaking pedagogical grammar’. It is seen by many as the most significant work of the early twentieth century in terms of asserting the importance of the inclusion of the spoken word in grammars of English (Carter and McCarthy 2017). Based on his observations of recorded spoken language, it shows an in-depth understanding of the grammar of speech and discourse. This includes insights on the greater use of parataxis and coordination of clauses rather than subordination in informal spoken contexts.

It is important to underline the salience of recording technology at this point in history. Though rudimentary, it allowed for the close study of phonetics and in so doing, Palmer (1924), among others, was also drawn to notice and detail how the grammar of spoken language manifests in different ways. This indeed was the harbinger of a whole new way of talking about grammar where the grammar writer moves from prescription and proscription, with a moralistic and aspirational perspective, to impartially describing what language users actually do with the grammar of a language when they use it in speaking and in writing. The advances in technology, from early sound recording equipment to the eventual availability of computers, the design of software and the ample storage capacity for data would ultimately render all grammars from a state of prescription to description, through the use of language corpora. As the next section will discuss, these digital advances, facilitated by corpus linguistics, fundamentally changed how we describe and research grammar and drove a new field of research.

The influence of corpora on the study of grammar

Hunston (2002) points out that corpus linguistics does not alter the nature of language, but does alter how we perceive it. The advent of language corpora profoundly changed the method of examining and describing grammar in English. As noted by Leech (2015), the emergence of corpora as empirical tools for evidencing and describing how grammar is used can be traced back to pre-corpus pre-digital grammarians who worked descriptively on grammar in mainland Europe in the early twentieth century. Grammarians such as Jespersen (1909–1949), Poutsma (1929) and Kruisinga (1909–1911/1931–1932), among others. As Leech (2015) explains, these grammarians worked manually from written notes of attestations and collections of language which they collected on slips of paper. The paper slips, which ‘could be annotated and shuffled at will, provided the nearest thing to the present-day corpus linguist’s “sort” button’ (Leech 2015: 147). Therefore, pre-technology, the shift from prescribing rules to describing what grammar was observed was already in train. The digital

turn brought an even greater empirical and objective lens to grammar and moved it further from the moralistic focus of previous centuries.

The work of Randolph Quirk and his associates is seen as the turning point in the use of digital means in the building on corpora as evidence bases for the large-scale description of grammar. The early corpus-based works, Quirk et al. (1972) and (1985), still stand as seminal exemplars of modern grammars of English. As Linn (2006: 86) notes in relation to Quirk et al. (1972), it ‘would prove to be a productive patriarch over the following decades’.

Two language collection projects (which pre-dated modern computing) led to the construction of the first electronic corpus, the London-Lund Corpus (LLC). These projects were the 1959 Survey of English Usage (SEU), a project of the University of London, and a sister project, the 1975 Survey of Spoken English (SSE), at the Lund University. Data from these projects informed Quirk et al. (1972) and Quirk et al. (1985). As detailed by Crystal and Quirk (1964), these started out as reel-to-reel recordings which were transcribed onto paper. The transcription process meant the development of an annotation system for grammatical structures as well as prosodic and paralinguistic features. The first copies of the computerised LLC were released in 1980. As Kauhanen (2011) details, the original LLC comprised 87 texts. The complete LLC stands at 100 texts. It is interesting to observe its structure and composition and to note the use of surreptitious recordings (see Table 9.1), which in current ethical protocols are no longer permissible. It is fair to note that in the early days of corpus linguistics, and digital recording in general, that such ethical concerns and data protection issues had not yet come to the fore.

More recent examples of grammars based on corpora had the luxury of substantially larger samples of language. As products of the Internet age, not only did corpora become substantially larger but their data were quicker to access and process due to the rapid developments in computer processing power and corpus software since the early work on the LLC. This era also saw an important link between research and ELT publishing. Pioneered by Professor John Sinclair in the 1980s, the partnership between Collins Publishers and the University of Birmingham (COBUILD) led to development of the COBUILD corpus, considered to be the first mega-corpus of its day. Its primary focus was lexical research.

Table 9.1 Breakdown of text types in the London-Lund Corpus

	<i>Text type</i>	<i>No. of texts</i>	<i>Word count</i>
S.1-S.3	Spontaneous, surreptitiously recorded conversations between intimates and distant	35	175,000
S.4	Conversations between intimates and equals	7	35,000
S.5	Non-surreptitious public and private conversations	13	65,000
S.6	Non-surreptitious conversations between dispartes	9	45,000
S.7	Surreptitious telephone conversations between personal friends	3	15,000
S.8	Surreptitious telephone conversations between business associates	4	20,000
S.9	Surreptitious telephone conversations between dispartes	5	25,000
S.10	Spontaneous commentary	11	55,000
S.11	Spontaneous oration	6	30,000
S.12	Prepared but unscripted oration	7	35,000
	Total	100	500,000

It revolutionised lexicography, allowing lexicographers to examine a wide range of real uses of a word and to identify word and pattern frequencies through concordancing software developed for this purpose. In 1987, this led to the publication of the ground-breaking COBUILD Dictionary. This set a new standard for how dictionaries were written. Subsequently COBUILD grammars followed, and again these grammar publications broke new ground in terms of empirically describing grammar and also in highlighting the importance that of collocational and colligational patterns. These patterns were identified as a result of corpus technology. At the time, the COBUILD Corpus was not publicly available but in 1991, the Bank of English was launched as a spin-off of the University of Birmingham Sinclair-inspired COBUILD project. The scale of the COBUILD project must be appreciated within the processing and storage limitations of the 1980s.

As research interest grew in the use of large-scale corpora, so did the focus on corpus design, and this ran in parallel with ever-increasing leaps in computer processing power and storage capacity. While having a large dataset was seen as a primary, there was an increasing attention given to how to represent a whole variety of English in a balanced way. A good example of this is the development of the British National Corpus, which was released in 1994. This brought together the might of technology at that time to allow for large-scale optical scanning of documents, the development of part-of-speech (POS) tagging and the use of audio-recording devices that were used to gather recordings from a representative sample of British speakers of English (see Crowdy 1993, 1994) to make up the 10 percent spoken component of the corpus. Through this corpus project, the concerted effort continued to push boundaries and created useful tools and protocols for corpus design, data collection, text typology, coding and POS tagging (Aston and Burnard 1998; Crowdy 1993, 1994).

The gathering and storing of data on a grand scale required funding and another innovation that benefitted from a partnership with a major publisher was *The Longman Grammar of Spoken and Written English* (Biber et al. 1999), which drew on the Longman Spoken and Written English (LSWE) corpus comprising more than 40 million words of data. The major innovation of these data was their categorisation across registers. The core registers of conversation, fiction, news and academic prose were represented in the LSWE and were spread across British and American English (see Biber et al. 1999: 24–35 for a detailed description of the corpus data). This allowed the authors not only to talk about grammar descriptively, based on this large dataset, but they were also able to describe how grammar varied across registers and language varieties. This register-calibrated corpus allowed Biber et al.’s (1999) grammar to give statistical results illustrating how language patterns varied across registers. For instance, when describing phrasal verbs (verb + adverbial particle, for example, *pick up*, *break down*, *point out*), they are able to provide more fine-grained details about their use across the core registers which are represented in their corpus. For instance, they can tell us that (Biber et al. 1999: 408–411):

- Phrasal verbs are used most commonly in fiction and conversation and they are relatively rare in academic prose
- In fiction and conversation, phrasal verbs occur almost 2000 times per million words
- *Go on* is the most common phrasal verb in the corpus

Grammatical description using a corpus was not always easy. As Leech (2015) notes, concomitant with the ability to look at large samples of language using a corpus came the responsibility (and sometimes burden) of unbiased objectivity through the ‘principle of total accountability’ (ibid: 146). That is, the grammarian was duty-bound to make use of all

relevant data in the corpus and not to shy away from data that does not appear to fit received paradigms. As Leech puts it, a grammarian using a corpus, ‘cannot ignore data that do not fit his/her view of grammar, as theoretically oriented grammarians have sometimes done in the past’ (2015: 147).

For those studying the ever-growing samples of spoken and written corpus data, it became clear that the grammar of spoken language did not always fit the written language paradigm. As detailed earlier, Biber et al. (1999) were able to show how conversation often had varying frequencies and functions in relation to many grammar forms. The publication of *The Cambridge Grammar of Spoken and Written English* (Carter and McCarthy 2006) was another marker of the impact of having the technology to record, transcribe and analyse spoken data alongside written corpus data. This work drew on the Cambridge International Corpus (CIC) as its resource (a database of almost 2 billion words at the time of writing). The major breakthrough of Carter and McCarthy’s (2006) work is the description of ‘spoken grammar’. The book’s two opening units are entitled ‘Spoken language’ and ‘Grammar and discourse’. Thereafter, it follows the established pathway through word classes, phrases and clauses, among other topics. The inclusion and foregrounding of the grammar of spoken language was a ground-breaking shift in grammar paradigms that now included details of the grammar of spoken language in its own right. As discussed earlier, the origins of spoken grammar can be traced back to early audio-recording technology used to gather spoken language by scholars such as Palmer (see Carter and McCarthy 2017). The link to technology is to be stressed here. As recording tools advanced and more and more spoken data was gathered, a greater body of evidence amassed to show that there were patterns of grammar in spoken language which were different than in writing. Many of the differences relate to the real-time nature of spoken language. As Carter and McCarthy (2006: 168) note:

In writing there are usually opportunities to plan and hierarchically structure the text. The writer can rephrase or edit what is written. In speech, utterances are linked together as if in a chain. One piece of information follows after another and speakers have few opportunities for starting again.

These real-time conditions led to a different structuring within spoken language, leading to observations such as (based on Carter and McCarthy 1995; Biber et al. 1999; Tao and McCarthy 2001; Carter and McCarthy 2006: 169–171; Rühlemann 2007; Timmis 2013; McCarthy and O’Keeffe 2014):

- Greater use of pronouns in spoken language, referencing the shared and immediate context of the speakers
- A tendency for simple noun phrases followed by a series of modifiers rather than full noun phrases with multiple pre-modifiers before the head noun
- The use of ‘phrasal chaining’ – where phrases are strung together like train carriages
- Increased use of ellipsis, especially situational ellipsis, as interlocutors share the same physical context
- More frequent use of subordination with or without the need for a dependent main clause, sometimes across turns, where one speaker will finish an interlocutor’s turn with a subordinate clause

Another important impact of corpus technology to note is that it allowed for consolidation and validation of existing smaller-scale findings that had emerged from other methodologies

such as critical discourse analysis (CDA) and conversation analysis (CA). These methodologies take a rigorous and in-depth approach to short texts or segments of spoken interactions, respectively, to generate findings about how language was used and constructed. Many researchers have applied CL approaches to build on CDA and CL findings and to test them on large corpus datasets. One area that has received much CL attention is that of turns construction. Systematicity in turn construction is at the core of CA (Sacks et al. 1974). For example, many CA works have looked at telephone call openings (Hutchby 1995; Thornborrow 2001) and established the systematic or ‘canonical’ order of turns. Corpus linguists have been able to use these findings as baselines to test on their data. For example, O’Keeffe (2006) examines call openings in a corpus of radio phone-ins, and she is able to draw on and compare her findings with CA results from work such as Whalen and Zimmerman (1987) on call openings in emergency response phone calls and Drew and Chilton (2000) on phone calls home from a daughter to her mother. CL findings both support and expand the existing insights from CA studies. Others have used corpora to look at how turns can be co-constructed in spoken language (Tao 2003; Rühlemann 2007; Tao and McCarthy 2001; Clancy and McCarthy 2015). Again, this is an area which has a broad base of research in the area of CA and where micro-findings from CA are enhanced and expanded through the use of corpus applications. Walsh et al. (2011) make a strong case for a closer alignment between CA and CL as a research methodology.

In this section we have looked at the impact of corpora, related technological advances and the advent of corpus linguistics on descriptive grammars. In summary, it is worth noting that while spoken corpora have resulted in a major paradigm shift in grammatical description, a great many spoken grammatical insights would not have been possible without comparisons with grammar statistics derived from written corpora (as Biber et al.’s. 1999 and Carter and McCarthy’s 2006 grammars amply demonstrate). We will now consider the specific issues and approaches related to the study of grammar items.

Critical issues and topics

How spoken data are transcribed, tagged and annotated to enhance the study of spoken grammar

A major innovation in the evolution of written corpus linguistic tools in terms of analysing grammar was automated POS tagging. Leech (2015: 148) points out that this innovation meant that it became relatively simple to extract instances of even ‘abstract categories such as progressive aspect or passive voice’. The next level up from POS tagging is to automatically parse a corpus syntactically, but as Leech puts it, this has ‘proved a more difficult nut to crack’ (Leech 2015: 149). This becomes even more complex when one moves further ‘up’ the discourse levels.

At a more fundamental level, as noted by Cook (1995), Kirk and Andersen (2016) and Kirk (2016), transcription of spoken data has proven very challenging, and this creates a more basic challenge to the study of spoken grammar and interaction. Kirk and Andersen (2016: 284) believe that ‘as currently conventionalised, transcriptions are not the spoken language; they amount to no more than selections of linguistic features from what through utterances was intersubjectively communicated’. Kirk and Andersen’s (2016) concern relates to the quest for a true encapsulation of spoken interaction in the written form of a transcript. Crucially, they note that understanding these deficiencies is a key to the ongoing development of pragmatic annotation: ‘The more linguists come to understand about those

interpersonal, intersubjective, communicative ways, the more new layers may be added to the linguistic structures which have been conventionally represented hitherto' (Kirk and Andersen 2016: 295).

To fully study an interaction and how language forms are functioning in a given set of contextual conditions, one would need a fuller form of pragmatically rich conventions for transcription. Kirk and Andersen (2016) survey some of the challenges of pragmatic annotation, including the fact that when real spoken language is transcribed, it is reduced into a pragmatically bereft form. What is not included in conventional lexico-syntactic transcription Kirk and Andersen (2016: 294–295) point out, are

indications of the pragmatics operating in an utterance: the illocutionary force or intent (the speech act status), the perlocutionary effect, the upholding or breaching of the Gricean co-operative principle, the politeness strategy invoked, the attitude of a speaker to the message of the utterance being made (pragmatic stance) or to the hearer of that utterance (face negotiation), and so to its potential impact. Much of what speakers utter is determined by a speaker's attitude towards what they are saying and towards the person(s) to whom they are saying it.

O'Keeffe et al. (2011) referred to pragmatically annotated corpora as the 'holy grail' for corpus pragmatic research; however, this seems increasingly possible. As summarised by Rühlemann and Aijmer (2015) there is a growing body of pragmatically annotated corpora, including speech acts (Stiles 1992; Garcia 2007; Kallen and Kirk 2012; Kirk 2016), discourse markers (Kallen and Kirk 2012; Kirk 2016), quotatives (Kallen and Kirk 2012; Kirk 2016; Rühlemann and O'Donnell 2012), participation role (Rühlemann and O'Donnell 2012) and politeness (Danescu-Niculescu-Mizil et al. 2013). To fully describe spoken grammar, these advances continue to be crucial.

How learner grammar is viewed: error or competency focus

Two pre-corpus strands of investigation into learner grammar were brought together in the 1990s by corpus linguistics. First, the 1940s saw the emergence of a contrastive focus on the impact of a learner's L1 on their learning of an L2. This led to the study of contrastive analysis of learner L1s as a means to predict what grammar (as well as vocabulary and phonology) might interfere in the learning of a given L2. The early work in this area was led by Robert Lado (see Lado 1957). It had a pedagogical goal, that is, to inform syllabi in what was a new approach to second language teaching at the time, namely the audiolingual approach. It was hoped that by anticipating the errors that a learner might make as a result of the grammatical differences between their first language and the target language, one could highlight and pre-empt these within the language teaching syllabus. This notion of error eradication was set within a behaviourist paradigm and was later superseded by a more cognitive approach to language learning and a more communicative approach to language teaching in the form of communicative language teaching (see Howatt and Widdowson 2004).

The other key area of learner grammar analysis that pre-dated corpora (and was more in line with a cognitive approach to learning, in contrast to Lado's work) was that of Pit Corder. Corder's (1967) work *The Significance of Learners' Errors* framed learner 'errors' not as 'bad habits' to be eradicated, but as a window into the learning process. This set the scene for the empirical study of learner grammar. The notion of 'interlanguage' coined by Selinker (1972) was a crucial milestone in the study of learner language as it gave it a name and independent

status to the cognitive process of learning a language where grammar 'errors' are a common developmental feature. Tono and Díez-Bedmar (2014: 164) point out that due in part 'to the technological limitations at the time (and the use of pre-computer learner corpora)', most of the learner data from this period was either discarded or shelved. The major turnaround in the 1990s came with the advent of contrastive interlanguage analysis (CIA) led by Sylviane Granger and subsequently the pioneering International Corpus of Learner English (ICLE) project (Granger 1994), allowing for not just for the large-scale study of interlanguage but also for contrastive analysis of L1s and L2s. The work of Granger therefore brought together and developed the two aforementioned antecedent strands of research into learner grammar: contrastive analysis of L1 versus L2 and error analysis of interlanguage. Granger's goal not just to look at learner language in contrast with a 'native' language but also to look at samples learner languages in comparison with each other, particularly from different L1s, in the broader framework of New Englishes. Her ICLE project was always linked to the broader International Corpus of English (ICE) (Greenbaum 1990) project, which aimed to gather both written and spoken samples of world Englishes. Through Granger's careful corpus design, she facilitated the comparison of learner English from different L1s but also the comparison of, for example, English from ICE Hong Kong Chinese speakers with Chinese L1 uses of learner (interlanguage) English (see Meunier et al. 2012).

Additionally, as the CIA field has developed, so has the almost exclusive LCR domination of focus on L2 English gradually become diluted. At the time of writing the 'Learner corpora around the world website'¹ lists 166 corpus projects, of which around 40 percent have a focus on over 20 L1s other than English. As Granger highlights (2015) while L1 English continues to dominate the field, recent years have seen a diversification in CIA studies in other languages (Zinsmeister and Breckle 2012; Jantunen and Brunni 2013, among others).

These CIA studies focused largely on argument writing, as this was a key skill for the types of learners that the field focused on – advanced learners in academic contexts (Granger 2015). The writing in the ICLE was carefully controlled by year of study and by the tasks that were undertaken to produce the writing samples. A detailed error coding system was developed as part of the process. This careful consideration of corpus design has been crucial to the growth of the ICLE and its value as a research resource internationally because it allows for comparison of output from different L1 backgrounds across year of study. Though writing continued to dominate CIA studies, there was an increased focus on learner speech, as evidenced in the spoken corpus the *Louvain International Database of Spoken English Interlanguage* (LINDSEI) (Gilquin et al. 2010), which was launched in 2011, and designed as the spoken counterpart for the ICLE. Other corpora led by Sylviane Granger include LOCNESS, PLECI and LONGDALE.

Contrastive analysis within corpus linguistics became a robust endeavour through the development of parallel and comparable corpora. As a result, systematic empirical work could be conducted on how two or more languages differed across lexis, syntax and beyond on a scale that had hitherto not been possible (Gilquin 2008). As noted by Gilquin (2008: 6), in reference to the impact of learner corpora such as ICLE, 'not only have such corpora allowed for the contextualisation of errors, but they have also enabled researchers to investigate what learners get right, what they overuse, i.e. use significantly more than native speakers, and what they under use, i.e. use significantly less than native speakers'.

Since the early 1990s, led and driven by the work of Sylviane Granger, a robust body of learner corpora, approaches to learner corpus design, error-coding and tagging systems have evolved. However, it is fair to say that corpus-based research on learner grammar has been focused on predicting, describing, comparing and quantifying learner errors. However, there

is a growing body of work that is focused on learners' grammatical competence. Looking at what learners can do is the flip side of grammatical errors, and this points to a new parallel approach emerging. To analyse learner competence, there is need to have as one's starting point learner data which is calibrated to some established marker of competence. The Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001) is increasingly offering such a calibration. The CEFR was developed at the turn of the century as a means of benchmarking language competence within Europe. The CEFR divides language competence levels into six stages (A1, A2, B1, B2, C1, C2), where A1 is the lowest and C2 is the highest. At each level, across a number of domains, the CEFR offers competency statements (referred to as 'can do' statements). Whereas hitherto, learner corpora have used 'year of study' as a proxy for the learner's level (for example, first-year university French students' English writing), the learners' CEFR level now offers an established marker of level. Increasingly, international exams (and pedagogical resources) are benchmarking to the CEFR, and therefore if a learner has taken an international English exam at a given level, it is possible to establish what CEFR level this correlates to.

This calibration of exam data to the CEFR is not without its critics and caveats. Carlsen (2012) notes that it can be arbitrary: 'levels of proficiency are not always carefully defined, and the claims about proficiency levels are seldom supported by empirical evidence' (2012: 162); see also O'Sullivan (2011: 265) who also warns of the potential arbitrariness of linking language test scores to the CEFR. The CEFR competency statements have also been criticised for being quite generic and 'underspecified' (see Milanovic 2009; Hawkins and Filipović 2012), as they were developed for all languages. Callies and Zaytseva (2013) claim that this has led to an increasing awareness of the need to develop more detailed linguistic descriptors.

The drive to empirically investigate what learners can do with grammar and vocabulary at different levels of the CEFR has driven new competency-based work on learner corpora. A recent large-scale corpus project, *The English Grammar Profile*, used the Cambridge Learner Corpus (a 55-million-word database of learner exam data from the Cambridge English exam suite) to develop a detailed profile or description of the grammar which learners know at each of the CEFR levels as evidenced from the Cambridge English exams that are linked to each level (www.englishprofile.org/english-grammar-profile/egp-online) (see O'Keeffe and Mark 2017 for a detailed description). The *English Grammar Profile* is a database of 1222 competency statements for learner grammar and is a major attempt to describe what learners can do at each level of the CEFR rather than what they cannot. It adds to other studies of learner grammar which also look at a smaller scale how learners' grammar develops in terms of not just what new forms learners can use at each level but also at what new ways they can use these forms (Harrison and Barker 2015; Hawkins and Buttery 2009, 2010; Hawkins and Filipović 2012; McCarthy 2016; Murakami and Alexopoulou 2016; Thewissen 2013).

Current contributions and research

Corpora, grammar and second language acquisition

The quest to establish how learners acquire a second language has been going on for many years but it is now increasingly aided by corpus linguistics. Psycholinguists working with corpus linguistics are now able to triangulate experimental methods in psycholinguistics using corpus data (Ellis et al. 2015; Simpson-Vlach and Ellis 2010; Durrant and Schmitt

2010; O’Donnell et al. 2013). By using corpus data, they have been able to increasingly substantiate usage-based theories of language which holds that ‘L2 learners acquire constructions from the abstraction of patterns of form-meaning correspondence in their usage experience’ (Ellis et al. 2015). Essentially, commonly used form-meaning pairings are understood to become ‘entrenched as grammatical knowledge in the speaker’s mind, the degree of entrenchment being proportional to the frequency of usage’ (Ellis and Ferreira-Junior 2009: 188).

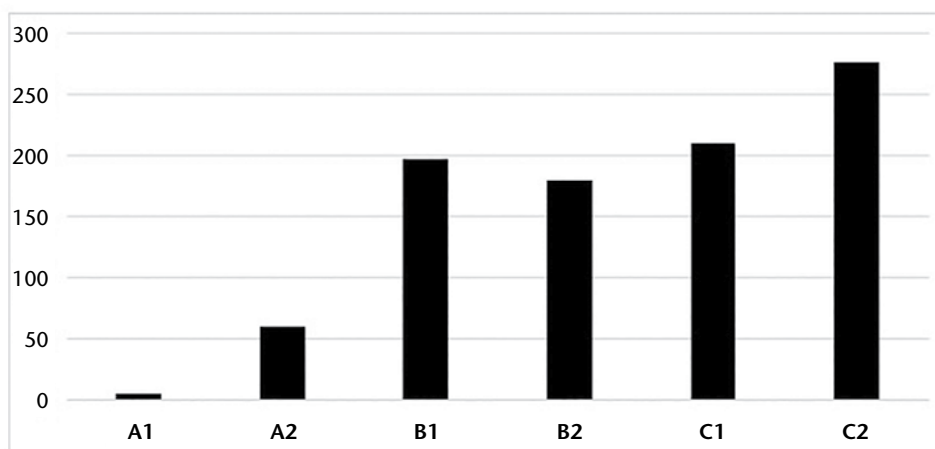
Ellis and Ferreira-Junior (2009) note the language which learners experience (i.e. their input) is linked directly to the patterns that they acquire. Creative linguistic competence comes, they say, from ‘the collaboration of the memories of all of the utterances in a learner’s entire history of language use and the frequency-biased abstraction of regularities within them’ (Ellis and Ferreira-Junior 2009: 188). Lately, learner corpora have been used to prove this theory and are ‘showing the evidence of learner formulaic use’. This is because corpora enable ‘the charting of the growth of learner usage’ (Ellis et al. 2015: 358). Ellis and his associates have been able to use large corpora such as the British National Corpus (BNC) (100 million words) or the Corpus of Contemporary American English (COCA) (560 million words) as a baseline of ‘typical language experience which serves learners as their evidence of learning’ (Ellis et al. 2015: 358).

The use of corpora in the empirical study of learner language has recently overturned some existing assumptions about the learning of grammar. A good example of this is the work of Murakami (2013) and (Murakami and Alexopoulou 2016). Until their in-depth work on learner corpus data, it had been held that there was a universal order in which grammar was learned. The long-standing paradigm was based on the pre-corpus work of Dulay and Burt (1973) who looked at the order of acquisition of eight grammatical morphemes. They studied the acquisition of these features across a cohort of 151 children who were learning English (all of whom had Spanish as their L1). Dulay and Burt (1973) study showed that all three cohorts followed the same order of acquisition of the eight grammar items under scrutiny and they concluded that L2 acquisition was guided by universal strategies. This was referred to as the ‘natural order’ of acquisition and it differed from the order identified previously by Brown (1973) who looked at L1 acquisition (Meunier 2015). Murakami’s (2013) study of the 55-million-word Cambridge Learner Corpus (CLC) revisited this long-held paradigm and he showed that, across a vastly larger data sample (including many L1 backgrounds), the notion of a natural or universal order of acquisition did not hold up and that the learner’s L1 had various influences on the order of morpheme acquisition (see Murakami 2013; Murakami and Alexopoulou 2016). This is a good example of how the advent of large-scale corpus research has allowed for greater depth and scope of research, and it has fundamentally changed our understanding of how grammar is acquired.

Sample analysis

We will now give a sample of how a corpus can be used to investigate grammar in a learner corpus. It will illustrate the use of a uniform corpus query language (CQL) can be used to readily retrieve data on learner and native speaker uses of a given grammar pattern. We will illustrate how this facilitates a more in-depth understanding of the developmental nature of how learners acquire and use grammar in terms of both form and function. Our data comprises a sample of 47,842,490 words from the CLC, which as detailed earlier, is made up of Cambridge English exam scripts and is calibrated to the six CEFR levels. We used the CQL

Table 9.2 Frequencies (per million words) of the modal verb *must* in affirmative and negative patterns in a sample of the CLC across the six levels of the CEFR



query [tag="PP|N.*"] [word="must"] [word="not|n't"] {0,1} [tag="VV"] to find all affirmative and negative instances of the modal verb *must*.

The results (as illustrated in Table 9.2) show a reasonably predictable cline of progression in terms of frequency of use of the form as learners progress from beginners' level to the most proficient level, C2. Interestingly, the average frequency of the affirmative and negative patterns of the modal verb *must* in the learner data is 131 occurrences per million words while the same CQL search string generates a frequency of just 5.4 occurrences per million words in the BNC written component. Examples from the English Grammar Profile, based on the CLC:

*You **must** wear your sports shoes and you must also bring your racket!*
[A1 Greek learner – English Grammar Profile]

*[talking about a guide book] In addition I **must** strongly recommend you add something about nightlife. (Russia; B2 VANTAGE; 2002; Russian; Pass)*
– [B2 Russian learner – English Grammar Profile]

Examples from the BNC for the same CQL search string include:

I have made my plans and I must stick to them – [Doc. Id. 589265]
Of course, we must not exaggerate the value of doubt and forget its darker side – [Doc. Id. 589612]

This tells us that the learners use this modal verb almost 25 times more in their exam tasks than one would find in a broad range of written texts in English. Considering that the learner data are from English language exam data, this is perhaps not too surprising, but what is important in terms of the value of having large corpora such as the BNC is that it can be used as a baseline for other research to bring in to relief the relative (in)frequency of a particular form in a specific context of use (in this case learner English exam data).

Table 9.3 A summary of the development of competencies in form (shaded) and function of affirmative and negative patterns of *must*, across forms and functions, taken from the English Grammar Profile (O'Keeffe and Mark 2017)

Competency	CEFR level					
	A1	A2	B1	B2	C1	C2
Form: Can use affirmative		✓	✓	✓	✓	✓
Form: Can use negative		✓	✓	✓	✓	✓
Function: Can use <i>must</i> to talk about obligation and necessity		✓	✓	✓	✓	✓
Form: Can use <i>must</i> with adverbs			✓	✓	✓	✓
Form: Can use ellipted forms of <i>must</i>			✓	✓	✓	✓
Function: Can use <i>must</i> to talk about deductions and conclusions			✓	✓	✓	✓
Function: Can use <i>must</i> to make invitations			✓	✓	✓	✓
Function: Can use <i>must</i> strong suggestions			✓	✓	✓	✓
Form: Can use <i>must have +ed</i>				✓	✓	✓
Function: Can use <i>must</i> to express concession (<i>I must admit</i>)				✓	✓	✓
Function: Can use <i>must</i> for emphasis				✓	✓	✓
Function: Can use <i>must</i> to express rules				✓	✓	✓
Form: Can use ellipted form <i>must have + -ed</i> (without subject)					✓	✓
Form: Can use inverted patterns of <i>must</i> after negative phrases (<i>not only, never, nor</i>) to bring focus or emphasis					✓	✓
Form: Can use past negative form						✓

An example of how this kind of research finding can have a wider impact and pedagogical application to English language education, we will briefly look at how this finding (along with many others from the CLC) were used to build the *English Grammar Profile* (EGP) resource. In the case of this finding on *must*, the results tell us that when the *must* patterns are broken down and looked at qualitatively, there is a growth across the competency levels and that this development is beyond frequency of use. On close analysis of the concordance

lines, it was found that the learners use the form in more sophisticated ways, as illustrated in Table 9.3 (see O’Keeffe and Mark 2017). While, on one hand, it reveals a grown in knowledge and use of the grammatical form, it also shows us that learners find new ways of using it.

The summary in Table 9.3 is a good illustration of the developmental pathway of a learner. While they learn the syntactic pattern at an early stage, it takes time to acquire more complex patterns. As they develop, they also acquire more functions in terms of how to use the forms that they have learned. A learner knows the syntactic form pronoun + modal verb + main verb at A2 but even at B2, we see that learners can use this pattern in a new way, namely to express concession, for instance, *However, I must admit that I completely agree with Chris* (example from the EGP).

This developmental form–function distinction is important because it can explain some of the mysteries of why learners seem to plateau at B1/B2 level. When researchers look at learner errors, they see a lack of progress in accuracy (i.e. error reduction) as learners go up levels of competency. For example, Thewissen (2013) illustrates lack of significant progress in accuracy between B2 and C2 levels. She points to the accuracy–complexity trade-off which learners make at higher levels when they take more risks and use more complex forms in more complex ways, but in doing so they increase the risk of making more errors. Analysis of grammatical errors counts alone reveals a plateau, suggesting a lack of development. In relation to this seeming lack of development between B2 and C2, Thewissen (2013: 87) points out that ‘it should not in and of itself be interpreted as a sign of no learning’. Thewissen sees this pattern as a ‘ceiling effect’ (after Milton and Meara 1995) and she argues that although errors still remain at these higher levels, there is a significant amount of learning taking place. In Larsen-Freeman’s words, this is not a case of ‘linguistic rigor mortis’ (2006: 597).

In our brief look at *must*, we have only focused on growing competency. This is the flip side of looking at error rates and yet it complements existing learner corpus work on errors. What is happening seems to be a double-edged. On one hand, learners acquire more forms and more ways of deploying these forms (functions) as they go up the levels. They make fewer errors overall, but errors still prevail because they are simultaneously trying out the new forms and functions in a pattern of growing competency. This supports Larsen-Freeman’s view that learner grammar is a dynamic system which, rather than being complete and acquired, is ‘never complete’ (2015: 496) and is in constant development. The finding also adds to the growing body of evidence that supports the usage-based view of both first and second language acquisition as discussed earlier (cf. Ellis et al. 2015). Pedagogically, it reinforces the importance of exposure to language forms both in receptive and productive processes and, most of all, it underscores the need to be more tolerant of errors in learner language because they evidence a cognitive process of linguistic risk-taking, which sometimes works and sometimes results in an error. However, as the patterns show, learners eventually get there. It just can look very messy while they are passing through B1/B2 levels!

The important point to glean from this sample analysis and other examples from research earlier is that our understanding of how grammar is learned is enhanced through the use of large datasets in the form of corpora. We are still at an early stage of our understanding of the process of grammar development but more and more corpus evidence is pointing to a strong link between grammar competency, risk (sometimes resulting in errors) and growing complexity.

Future directions

Corpora as digital databases have and continue to offer enhanced understandings of grammar through empirical studies. Even at the outset, as detailed earlier, they have led to better descriptions of language as it is used, and this has led to a move away from prescription of rules. Looking at large samples of language also brought to light the internal variation in grammar use across registers, as detailed in relation to the work of Biber et al. 1999, for instance. And over the years as the technology for gathering spoken data has improved, so has the availability of more spoken data and the paradigmatic shift in naming spoken grammar as separate to written grammar as detailed in relation to Carter and McCarthy (2006).

In terms of learner grammar and the study of interlanguage, there is scope for major growth in this area with the emergence of studies based on corpus data which are stratified to competency benchmarks. The move to look at grammar competency will also offer a fuller understanding of why many learners seem to reach a syntactic ceiling in their developmental path. This additional focus on learner competence represents a welcome addition to complement the study of learner error. The availability of more longitudinal learner data, where individual learners are tracked over their learning journey, from level to level would be very valuable data, not least of all in highlighting the variability between individual learners (see Gablasova et al. 2017).

As we noted at the outset, technology and the study of texts have been well matched and so, as we have shown, DH has thrived in this respect. However, in the study of how languages are learner or acquired, the application of technology is at an early stage and a greater use of corpora in the investigation of second language acquisition would greatly enhance our quest to understand how grammar is or is not learned. Gablasova et al. (2017: 2) note that ‘information from L2 production can uncover linguistic patterns that point to underlying factors in SLA’. As Myles (2015) points out, learner corpus research is very suitable to some of the main concerns of SLA, including the linguistic system underlying learners’ performance and how it is constructed at various stages of development in terms of phonology, morphology, lexis, syntax, semantics, discourse, pragmatics and the interface between these systems and the role of the L1, L2 and universal formal properties of human languages in shaping and/facilitating the stages of development of linguistic features. SLA research methodology needs to ‘move into the digital age’, Myles (2015: 309) advises, and learner corpus research needs to ‘take full account of developments in SLA theorising’.

Corpora will continue to be built and used to investigate and describe how we use grammatical patterns (among many other things) (McCarthy 2017b). As corpora stack up over time, they will have a new use in helping us describe how language is changing. Historical corpus linguistics is already a vibrant field where historical sources have been used in digital form to trace language development and change but as second, third and fourth generation replications of existing corpora come about, they will offer a robust like-with-like comparison of grammar patterns and this will most likely lead to very interesting insights into how grammar is changing, or not.

Note

- 1 <https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html>

Further reading

- 1 Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Pearson and Longman.

- 2 Carter, R.A. and McCarthy, M.J. (2006). *Cambridge grammar of English: A comprehensive guide to spoken and written grammar and usage*. Cambridge: Cambridge University Press.
(A truly comprehensive and accessible description of spoken and written grammar.)
- 3 Carter, R.A. and McCarthy, M.J. (2017). Spoken grammar: Where are we and where are we going? *Applied Linguistics* 38(1): 1–20.
(A reflection on 20 years of corpus-based work on spoken grammar, underlining the importance of its inclusion in any pedagogical grammar. The paper also offers an interesting summary of the history of grammars.)
- 4 Crystal, D. (2017). *Making sense: The glamorous story of English grammar*. Oxford: Oxford University Press.
(Crystal charts the development of grammar, its history, varieties irregularities and uses.)
- 5 Gablasova, D., Brezina, V. and McEnery, T. (2017). Exploring learner language through corpora: comparing and interpreting corpus frequency information. (A critical analysis of comparative learner corpus analysis.) *Language Learning* 67(S1): 130–154.
- 6 Linn, A.R. (2006). English grammar writing. In A. McMahon and B. Aarts (eds.), *The handbook of English linguistics*. Oxford: Blackwell, pp. 72–92.
(This chapter provides a comprehensive history of grammars, charting historical prescriptive grammars to modern corpus-based descriptive works.)

References

- Aitchison, J. (2001). *Language change: Progress or decay?* Cambridge: Cambridge University Press.
- Aston, G. and Burnard, L. (1998). *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh: Capstone.
- Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Pearson and Longman.
- Brown, R. (1973). *A first language: The early stages*. London: George Allen and Unwin.
- Bullokar, W. (1586). *Pamphlet for grammar*. Imprinted at London: By Edmund Bollifant.
- Callies, M. and Zaytseva, E. (2013). The corpus of academic learner English (CALE) – A new resource for the study and assessment of advanced language proficiency. In S. Granger, G. Gilquin and F. Meunier (eds.), *Twenty years of learner corpus research: Looking back, moving ahead. Corpora and language in use – proceedings 1*. Louvain-la-Neuve: Presses Universitaires de Louvain, pp. 49–59.
- Carter, R. and McCarthy, M. (1995). Grammar and the spoken language. *Applied Linguistics* 16(2): 141–158.
- Carter, R.A. and McCarthy, M.J. (2006). *Cambridge grammar of English: A comprehensive guide to spoken and written grammar and usage*. Cambridge: Cambridge University Press.
- Carter, R.A. and McCarthy, M.J. (2017). Spoken grammar: Where are we and where are we going? *Applied Linguistics* 38(1): 1–20.
- Carlsen, C. (2012). Proficiency level – a fuzzy variable in computer learner corpora. *Applied Linguistics* 3(2): 161–183.
- Clancy, B. and McCarthy, M. (2015). Co-constructed turn-taking. In book K. Aijmer and C. Ruhlmann (eds.), *Corpus pragmatics: A handbook*. Cambridge: Cambridge University Press, pp. 430–453.
- Cook, G. (1995). Theoretical issues: Transcribing the untranscribable. In G. Leech, G. Myers and J. Thomas (eds.), *Spoken English on computer: Transcription, mark-up and application*. Abingdon: Longman, pp. 35–53.
- Corder, S.P. (1967). The significance of learner's errors. *International Review of Applied Linguistics* 5: 161–169.
- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing* 8(4): 259–265.

- Crowdy, S. (1994). Spoken corpus transcription. *Literary and Linguistic Computing* 9(1): 25–28.
- Crystal, D. and Quirk, R. (1964). *Systems of prosodic and paralinguistic features in English* (No. 39). Berlin: Walter De Gruyter Inc.
- Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J. and Potts, C. (2013). *A computational approach to politeness with application to social factors*. In *Proceedings of ACL*. Available at: www.mpi-sws.org/~cristian/Politeness.html (Accessed January 2017).
- Drew, P. and Chilton, K. (2000). Calling just to keep in touch: Regular and habitual telephone calls as an environment for small talk. In J. Coupland (ed.), *Small talk*. London: Longman, pp. 137–162.
- Dulay, H.C. and Burt, M.K. (1973). Should we teach children syntax? *Language Learning* 23(2): 245–258.
- Durrant, P. and Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research* 26(2): 163–188.
- Ellis, N.C. and Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7: 188–221.
- Ellis, N.C., O'Donnell, M. and Römer, U. (2015). Usage-based language learning. In B. MacWhinney and W. O'Grady (eds.), *The handbook of language emergence*. Hoboken, NJ: Wiley, pp. 163–180.
- Gablasova, D., Brezina, V. and McEnery, T. (2017). Exploring learner language through corpora: Comparing and interpreting corpus frequency information. (A critical analysis of comparative learner corpus analysis.) *Language Learning* 67(S1): 130–154.
- Garcia, P. (2007). Pragmatics in academic contexts: A spoken corpus study. In M.C. Campoy and M.J. Luzón (eds.), *Spoken corpora in applied linguistics*. Bern: Peter Lang, pp. 97–128.
- Gilquin, G. (2008). Combining contrastive analysis and interlanguage analysis to apprehend transfer: Detection, explanation, evaluation. In G. Gilquin, S. Papp and M. Belén Díez-Bedmar (eds.), *Linking up contrastive and learner corpus research*. Amsterdam: Rodopi, pp. 3–33.
- Gilquin, G., De Cock, S. and Granger, S. (eds.) (2010). *Louvain international database of spoken English interlanguage*. Louvain: Presses universitaires de Louvain.
- Granger, S. (1994). The learner corpus: A revolution in applied linguistics. *English Today* 10: 25–33.
- Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1(1): 7–24.
- Greenbaum, S. (1990). Standard English and the international corpus of English. *World Englishes* 9: 79–83.
- Harrison, J. and Barker, F. (eds.) (2015). *English profile in practice. English profile studies, 5*. Cambridge: Cambridge University Press.
- Hawkins, J. A. & Buttery, P. (2009). Using learner language from corpora to profile levels of proficiency: Insights from the English Profile Programme. In Taylor, L. & Weir, C. J. (Eds). *Language Testing Matters: Investigating the Wider Social and Educational Impact of Assessment*, 158–175. Cambridge: Cambridge University Press.
- Hawkins, J. and Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal* 1: E5. doi:10.1017/S2041536210000103
- Hawkins, J. and Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the common European framework*. Cambridge: Cambridge University Press.
- Howatt, A.P.R. and Widdowson, H. (2004). *A history of English language teaching*. Oxford: Oxford University Press.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Hutchby, I. (1995). Aspects of recipient design in expert advice-giving on call-in radio. *Discourse Processes* 19: 219–238.
- Jantunen, J.H. and Brunn, S. (2013). Morphology, lexical priming and second language acquisition: A corpus-study on learner Finnish. *Twenty years of learner corpus research: Looking back, moving ahead*. Louvain-la-Neuve: Presses universitaires de Louvain, pp. 235–245.
- Jespersen, O. (1909–49). *A modern English grammar on historical principles* (7 vols). Copenhagen: E Munksgaard and London: Allen & Unwin.
- Kallen, J.L. and Kirk, J.M. (2012). *SPICE-Ireland: A user's guide*. Belfast: Cló Ollscoil na Banríona.

- Kauhanen, H. (2011). *The London-Lund corpus of spoken English*. Available at: www.helsinki.fi/varieng/CoRD/corpora/LLC/basic.html. Last updated 2011-03-21 by Henri Kauhanen.
- Kirk, J.M. (2016). The pragmatic annotation scheme of the SPICE-Ireland corpus. *International Journal of Pragmatics* 21(3): 299–322.
- Kirk, J.M. and Andersen, G. (2016). Compilation, transcription, markup and annotation of spoken corpora. *International Journal of Corpus Linguistics* 21(3): 291–298.
- Kirschenbaum, M.G. (2010). What is digital humanities and what's it doing in English Departments? *ADE Bulletin* 150: 55–61.
- Kruisinga, E. (1909–1911/1931–1932). *A handbook or present day English*. Groningen: Noordhoff.
- Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers: With a foreword by Charles C. Fries*. Ann Arbor: University of Michigan Press.
- Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics* 27: 590–619.
- Larsen-Freeman, D. (2014). Teaching grammar. In M. Celce-Murcia, D.M. Brinton and M.A. Snow (eds.), *Teaching English as a second or foreign language* (4th ed.). Boston, MA: National Geographic Learning, pp. 255–270.
- Larsen-Freeman, D. (2015). Saying what we mean: Making a case for 'language acquisition' to become 'language development'. *Language Teaching* 48(4): 491–505.
- Leech, G. (2015). Descriptive grammar. In D. Biber and R. Reppen (eds.), *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press, pp. 146–176.
- Linn, A.R. (2006). English grammar writing. In A. McMahon and B. Aarts (eds.), *The handbook of English linguistics*. Oxford: Blackwell, pp. 72–92.
- Lowth, R. (1762). *A short introduction to English grammar*. Fac. Rpt. 1967.
- McCarthy, M.J. (2016). Advanced level. In Hinkel, E. (ed.), *Teaching English grammar to speakers of other languages*. New York: Routledge.
- McCarthy, M.J. (2017a). *English grammar: Your questions answered*. Cambridge: ProLingua Publishing.
- McCarthy, M.J. (2017b). Usage on the move: Evolution and re-evolution. *Training, Language and Culture* 1: 8–21. doi:10.29366/2017tlc.1.2.1
- McCarthy, M.J. and O'Keeffe, A. (2010). Historical perspective: What are corpora and how have they evolved? In A. O'Keeffe and M. McCarthy (eds.), *The Routledge handbook of corpus linguistics*. London: Routledge, pp. 3–13.
- McCarthy, M.J. and O'Keeffe, A. (2014). Spoken grammar. In M. Celce-Murcia, D.M. Brinton and M.A. Snow (eds.), *Teaching English as a second or foreign language* (4th ed.). Boston, MA: National Geographic Learning, pp. 271–287.
- Meunier, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin and F. Meunier (eds.), *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press.
- Meunier, F., de Kock, S., Gilquin, G. and Paquot, M. (eds.) (2012). *A taste of corpora: In honour of Sylviane Granger*. Amsterdam: John Benjamins.
- Milanovic, M. (2009). Cambridge ESOL and the CEFR. *University of Cambridge ESOL Examinations Research Notes* 37: 2–5. Available at: www.cambridgeenglish.org/images/23156-research-notes-37.pdf.
- Milton, J. and Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL Review of Applied Linguistics* 107–108: 17–34.
- Murakami, A. (2013). Cross-linguistic influence on the accuracy order of L2 English grammatical morphemes. *Twenty Years of learner Corpus Research: Looking Back, Moving Ahead: Corpora and language in Use* 1: 325–334.
- Murakami, A. and Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes. *Studies in Second Language Acquisition* 38(3): 365–401.
- Murray, L. (1795). *English grammar: Adapted to the different classes of learners: With an appendix containing rules and observations for promoting perspicuity in speaking and writing*. York: Wilson, Spence, and Mawman.

- Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin and F. Meunier (eds.), *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press.
- O’Donnell, M.B., Römer, U. and Ellis, N.C. (2013). The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics* 18(1): 83–108.
- O’Keeffe, A. (2006). *Investigating media discourse*. Abingdon: Routledge.
- O’Keeffe, A., Clancy, B. and Adolphs, S. (2011). *Introducing pragmatics in use*. Abingdon: Routledge.
- O’Keeffe, A. and Mark, G. (2017). The English grammar profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics* 22(4): 457–489.
- O’Sullivan, B. (2011). Language testing. In J. Simpson (ed.), *The Routledge handbook of applied linguistics*. London: Routledge, pp. 259–273.
- Palmer, H.E. (1924/1969). *A grammar of spoken English on a strictly phonetic basis*. First edition 1924 Third edition 1969, revised and rewritten by R. Kingdon. Heffer.
- Poutsma, H. (1929). *A grammar of late modern English* (4 vols). Groningen: Noordhoff.
- Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1972). *A grammar of contemporary English*. London: Longman.
- Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.
- Rühlemann, C. (2007). *Conversation in context. A corpus- driven approach*. London and New York: Continuum.
- Rühlemann, C. and Aijmer, K. (2015). Corpus pragmatics: Laying the foundations. In K. Aijmer and C. Rühlemann (eds.), *Corpus pragmatics: A handbook*. Cambridge: Cambridge University Press, pp. 1–26.
- Rühlemann, C. and O’Donnell, M.B. (2012). Introducing a corpus of conversational narratives: Construction and annotation of the Narrative Corpus. *Corpus Linguistics and Linguistic Theory* 8(2): 313–350.
- Sacks, H., Schegloff, E.A. and Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Language* 50(4): 696–735.
- Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics* 10: 209–241.
- Shelley, E. (1848). *The people’s grammar; or English grammar without difficulties for ‘the million’*. Huddersfield: Bond and Hardy.
- Simpson-Vlach, R. and Ellis, N.C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31(4): 487–512.
- Smith, R.C. (1990). *The writings of Harold E. Palmer: An overview*. Tokyo: Hon-no-Tomosha Publishers.
- Stiles, W.B. (1992). *Describing talk: A taxonomy of verbal response modes*. Newbury Park: Sage Publications.
- Swan, M. (2011). Grammar. In J. Simpson (ed.), *The Routledge handbook of applied linguistics*. Taylor & Francis, pp. 557–570.
- Sweet, H. (1899). *The practical study of languages: A guide for teachers and learners*. London: J.M. Dent and Co.
- Sweet, H. (1900). *A new English grammar, logical and historical*. Oxford: Oxford University Press.
- Terras, M. (2016). A decade of digital humanities. *Journal of Siberian Federal University* 7: 1637–1650.
- Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error tagged learner corpus. *The Modern Language Journal* 97: 77–101.
- Timmis, I.G. (2013). Spoken language research: An applied linguistic challenge. In B. Tomlinson (ed.), *Applied linguistics and materials development*. London: Continuum, Bloomsbury, pp. 79–95.
- Tao, H. (2003). Turn initiators in spoken English: A corpus-based approach to interaction and grammar. *Language and Computers* 46: 187–207.
- Tao, H. and McCarthy, M.J. (2001). Understanding non-restrictive which-clauses in spoken English, which is not an easy thing. *Language Sciences* 23: 651–677.

- Thornborrow, J. (2001). Questions, control and the organisation of talk in calls to a radio phone-in. *Discourse Studies* 3(1): 119–143.
- Tono, Y. and Díez-Bedmar, M.B. (2014). Focus on learner writing at the beginning and intermediate stages: The ICCI corpus. *International Journal of Corpus Linguistics* 19(2): 163–177.
- Walsh, S., Morton, T. and O’Keeffe, A. (2011). Analysing university spoken interaction: A CL/CA approach. *International Journal of Corpus Linguistics* 16(3): 325–345.
- Whalen, M.R. and Zimmerman, D.H. (1987). Sequential and institutional context in calls for help. *Social Psychology Quarterly* 50(2): 172–185.
- Zinsmeister, H. and Breckle, M. (2012). The ALeSKo learner corpus. *Multilingual Corpora and Multilingual Corpus Analysis* 14: 71–96.

Introduction

The lexicon of a language is the stock of words on which speakers of that language can draw, which makes lexis arguably the most important factor in the study of any language. Linguists of every stripe can make a case for their own area of study to be viewed as the fundamental level of linguistics – phonologists argue for the primacy of sound, syntacticians make the point that structure is crucial to language and semanticists are clear that the purpose of communication is meaning – but if we were to ask language users what language is made up of, they would unhesitatingly reply *words* and would mean by this something like the mass of the lexicon as represented in a dictionary. This is understandable; when someone asks how large a language is, we naturally try to answer by counting words, we index words in the final pages of non-fiction and search for words when we hunt through digital sources and we also have more reference sources on words than we do for any other linguistic unit. Here, we therefore focus on research into lexis with regards to its status as a core concern for the digital study of language.

Lexis is important in the digital humanities (DH) not just because of all the work which goes into the study of words but also because lexical items are used to *represent* much in digital humanities research. DH has often been criticised as being dominated by text above all other forms of data – see, for example, Champion 2017 for an overview – and this is a fair statement, if perhaps one which conflates history and evolution with some sort of designer’s intent. Text has been ‘center stage’ within DH for some time (Hockey 2004), and with its ease of access, small file size and wide toolset availability, it is likely to continue to be there for some time to come. If text stands at centre stage for DH, then as we discuss later, words are centre stage for text.

However, very little work in DH is interested in words for their own sake – instead, lexis is used to measure a huge range of phenomena, including historical attitudes (e.g. Baker and McEnery 2017); cultural significance (Alexander and Struan 2013); legal meaning (Solan and Gales 2016 or Mouritsen 2010); an author’s style (Mahlberg 2013); the ‘aboutness’ of a text or collection of texts (Bondi and Scott 2010); and how much attention is being paid to something within a text, society, culture or period (Michel et al. 2010). This chapter

outlines these uses of lexis in DH, discusses certain critical issues and major resources with a particular but not exclusive focus on the intersection of lexis with semantics research using digital humanities techniques and illustrates the use of these by a case study of the English semantic field of *money*.

Background

Words, lexemes, types and tokens

Linguists distinguish between *words* and *lexemes*: a lexeme is an abstract unit which entails the set of possible forms a word can take. A lexeme is referred to using the conventional ‘dictionary form’ of a word (e.g. [to] *be*, *child*, *table*, or *close*) and includes all its various inflected forms (*be*, *am*, *are*, *is*, *was*, *were*, *being*, *been*; *child*, *children*; *table*, *tables*; *close*, *closer*, *closest*).

As an illustration of the value of the word/lexeme distinction, consider the following passage from the short story *Reginald on Besetting Sins*:

On a raw Wednesday morning, in a few ill-chosen words, she told the cook that she drank. She remembered the scene afterwards as vividly as though it had been painted in her mind by Abbey. The cook was a good cook, as cooks go; and as cooks go she went.
(Saki 2016: 32)

The final sentence, an example of parapsydokian, contains 15 words in total. However, *cook* is present four times (twice in the plural as *cooks*), *go* three times (the third in the past tense as *went*) and *as* twice. This means that the 15-word sentence contains only nine lexemes: *the*, *cook*, *be*, *go*, *a*, *good*, *as*, *and*, *she*.

A related distinction found in corpus studies is between *tokens* and *types*, using terminology from metaphysics. A token is an instance of type in use; in ‘the cat sat on the mat’ there are six tokens but five types; *the* is repeated, and thus constitutes two tokens of a single type. The ratio of types to tokens can be used as a measure of vocabulary richness (see Tweedie and Baayen 1998 for more on these measures and their problems).

Finally, one issue which arises at the confluence of the word/lexeme and type/token distinctions is that some software and researchers take *types* to mean *orthographic types* (that is, they count instances of a type as simply any duplicate string of letters of that type), and at other times texts are lemmatised (see later) to find *lexical types* (which are therefore lexemes). In the 15-token closing sentence of the Saki passage earlier, there are nine lexemes (and so nine types in the lexical sense) but 11 types in the orthographic sense of strings of letters (*the*, *cook*, *was*, *a*, *good*, *as*, *cooks*, *go*, *and*, *she*, *went*). Software or corpus processing tools which are not programmed to recognise that, for example, *went* is a form of *go* and the noun *cooks* is a derived form of *cook* are limited to an ability to identify that the letter string *cook* appears here twice and that *cooks* also appears twice, without acknowledging these as inflected forms of the same underlying lexeme. An additional distinction can be made between different parts of speech, if the corpus is tagged in that way, meaning that, in these instances, types consist of orthographic words within distinct grammatical categories. Such categorisation would typically identify *cooked* as a past tense verb form and *cooking* as present participle (or gerundival) form, without necessarily lemmatising them under the *cook* lexeme.

The analysis of lexis in digital text studies is complicated by items such as phrasal verbs (e.g. *slow down*, *bring up*) and expressions comprising multiple words (e.g. *upside down*, *top hat*, *Melton Mowbray*). These defy the lay understanding of words as strings of letters bookended by white space, and arguably are so closely semantically linked that it is inadvisable to treat them as separable – for example, although the phrase *buttoned down* does have something to say about the history and connotations of types of clothing, it is debatable whether its inclusion in a search for *button* or *buttoned* would be instructive for a researcher. Additionally, expressions such as these may not act as indivisible strings of text; ‘I’ll *shoot down* whatever argument you make’ and ‘I’ll *shoot* your argument *down*’ contain the same phrasal verb, but in the latter a noun phrase is interposed between *shoot* and *down*. Problems of this type are acknowledged by the increasing use of terms such as ‘lexical item’ to include expressions formed of multiple words and other word combinations which have strong arguments for coherence (for further discussion see later, and Moon 2015).

Words as proxies

When dealing with textual data, DH generally affirms the primacy of the word as the *unit* of analysis, as argued earlier. It should be said, though, that this is not generally because of any special interest in lexicology, but rather because of the overwhelming utility of words as *proxies* for other phenomena. For example, the raw frequency of a word within a text provides limited linguistic information, but when the frequency of a word is used as a proxy for how interested a text (or writer, or culture) is in the concept to which that word refers, the frequency becomes highly relevant to researchers. In the majority of DH research outwith lexicology, therefore, words are almost always used as some sort of proxy for other information, and when an analyst searches or browses a text by word, much of the skill involved in their research will be focused on finding good, robust and reliable proxies.

By way of example, where a researcher may wish to understand the representation of the concept of happiness in a text, a search for just the word *happiness* itself is unlikely to return the full results sought; the concept happiness has a variety of expressions depending on personal preference and nuances of the condition which the author is trying to convey. The researcher is therefore likely to be more interested in the concept as expressed using a number of words with semantic overlap, which may include *joy*, *delight* or *merriment*, so that the word *happiness* itself is a proxy used to express an idea with many lexical representations (although then care has to be taken only to find *delight* where it means happiness and not count instances, for example, of the sweet *Turkish delight*). Conversely, research in fields such as stylometry and authorship analysis are less interested in the usage of a particular word such as *happiness* than they are in whether grammatical and functional words of the text, and perhaps the accumulation of types of lexis, such as adjectives or adverbs, are indicative of the writing style adopted by an author. This type of textual analysis uses the words chosen by the author as a proxy for patterns in their use of language below the level of deliberate awareness.

Overall, the use of lexis as a proxy is not often explicitly taught, and while very often instinctively done well, at times it requires more conscious attention. A researcher must ask themselves the fundamental question *am I looking for what I need?* A failure to do so, and to find out if there are other ways of identifying such proxies, can mean that an attempt to answer big questions about cultural significance, authorial attention, structural opposition, authorship or other interesting applied linguistic objects of study, runs the risk of being too

narrowly focused on an orthographic search string which either bears little relation to or omits a sizeable proportion of the information which the researcher wishes to obtain. It is worth paying close attention to this issue.

Critical issues and topics

The approach to analysing text in the DH has undergone a revolution in methods and technology as both the quantity of data available and the processing power accessible to the average researcher have increased rapidly within a couple of decades. At the same time, many of the fundamental theoretical issues in lexical research have remained the same but are ever open to re-evaluation as the discipline develops. We organise this section around three themes which are informed by the challenges of lexical theory and recent progress in the DH: first, what do we count as words; second, what are the inherent challenges to studies of lexis; and finally, what problems do we face with using any sort of lexical data in English language research?

What do we count as words?

It is common to begin writing about lexis by struggling with a definition. People can quite easily intuitively grasp what a lexical item is; however, on further analysis the category proves to be fuzzy with no clear boundaries, which poses a challenge for lucid descriptions. Words in the English language tend to represent a unified semantic concept (Taylor 2015: 9–10), are often indivisible, belong to a particular grammatical class, move syntactically as a unit and when spoken aloud have a solitary primary stress. We normally also mark such units with spaces on either side in writing, excluding punctuation, and so this space-surrounded aspect of words is sometimes called the *orthographic criterion*. The word *table* – for the sake of example in the sense of a flat-topped piece of furniture, although the following is equally true for other senses – easily meets these criteria: it is a coherent concept, can't be interrupted (no **ta-wood-ble*), is a noun, moves around a sentence as a whole ('put it on the table', 'which table?', 'the table over there', etc.), has primary stress on the first syllable and indeed is written with a space before and after it. (See further Plag 2003: 4 ff.; Durkin 2009: 34 ff.)

However, any and all parts of that definition can be challenged. Counterexamples are easy to find, especially on the question of written spaces:

- Someone can *act up*, a unit with spaces inside but acting as a single verb and having an idiomatic meaning which relies on both parts.
- More than one *poet laureate* are *poets laureate*, both with internal spaces but also with the plural marker *-s* interrupting the unit.
- Someone can *spill the beans* meaning reveal previously secret information, but that idiom must be moved as a unit, so there is a strangeness about **the beans Helen spilled* intended in an idiomatic sense.
- The marshy ghost light *will-o'-the-wisp* can just as easily be written *will o' the wisp*, with the lack of spaces a reflection of personal preference or a publisher's house style, but not to be read as somehow transforming the word's lexical status.
- The title of the musical and movie *My Fair Lady* is recognisably a determiner–adjective–noun structure on one level but has a coherent semantic reference only together, and so acts as if the whole phrase were a single noun. If one were to put it in the plural – in

the context, say, of a DVD storehouse – the irregular plural of *lady* poses a significant challenge, especially in its written form; ‘go get 20 *My Fair Ladies*’ hypercorrects and seems to refer to 20 of a different movie, *My Fair Ladies*, while ‘go get 20 *My Fair Ladys*’ is more correct, as the irregular plural belongs to *lady* on its own rather than the *My Fair Lady* unit, but looks strange to many people, accustomed as they are to the irregular form *ladies*. Most English-speakers would, in all probability, use this expression in speech but avoid writing it down because both renderings are troublesome, and so a paraphrase such as ‘go get 20 copies of *My Fair Lady*’ would be less jarring.

Such questions are fascinating to lexicologists but can seem rather abstruse and are generally skimmed over in introductory linguistics courses. They come sharply into focus, however, when considering the nature of what we will call the *digital word* – the unit which in the digital humanities is searched for, defined and used as a proxy for all the things we outline earlier. In almost every case, words are treated digitally as if they are orthographic words: units separated by spaces or punctuation – not only *table*, but also *isn’*, and *t*; *act* and *up*; *My*, *Fair*, and *Ladys*; *will*, *o’*, *the*, and *wisp*; and the forward upper deck of a ship, *fo’c’s’le*, as a nightmare of individual letters and syllables. Researchers can tweak some of these rules (‘do not break words at apostrophes’ is a rule which some text-processing software can handle), but this is still problematic for a collection of texts which vary in house style, with some preferring, say, *will-o’-the-wisp* and others *will o’ the wisp*.

A solution which is often used is to bind appropriate orthographic words into multi-word expressions (MWEs), a concept derived from linguistic studies of phraseology (Weinreich 1969; see also Wray 2002 for a schematisation of these). In computational studies, these have been roughly defined as ‘interpretations that cross word boundaries (or spaces)’ (Sag et al. 2002: 2) and can be thought of as units which ‘override’ the orthographic criterion. In semantic analysis, decomposing idioms risks losing sight of their original referent concept, and so slightly under a third of the entries in the USAS semantic tagging lexicon are MWEs (Piao et al. 2003: 3; for more details of USAS see later). Sag et al. argue that the number of MWEs in a speaker’s lexicon is likely to be even larger than the number of single words, pointing out that a study of WordNet 1.7 revealed that 41 percent of its vocabulary consisted of MWEs and that this is likely to fall short of the actual number of such expressions which exist in the language (2002: 2). In practice, those combinations of orthographic words found in a list of MWEs would ‘override’ any consideration by the software of these words as separate items. Such MWEs often include idioms, such as *raining cats and dogs*, *once in a blue moon* and *bite the bullet*, which all have a unified meaning and do not actively evoke for native speakers the embedded concepts of cats, dogs, moons, bullets and so on (Gibbs 1985); it would be problematic to say that dogs (as animals) are frequently discussed in a text which habitually talks about rain using that idiom, although true to say that the lexeme appears. MWEs can also include names such as *New York City* or *My Fair Lady* which are similarly unified expressions and pose the same issues as idioms (there is nothing necessarily new about *New York*, nor mine about *My Fair Lady*; the human being called *Calvin Klein* is referred to in English far less frequently than his eponymous clothing, underwear or fragrances).

It is difficult to see, however, where this category ends. For example, definitions of MWEs can, and sometimes do, include other categories which could theoretically and usefully be incorporated. Clichés, for example, such as *name*, *rank* and *serial number* exist as unified phrases which are compositional in structure, which is to say that, rather than being idiomatic, their meaning is made up from the meanings of the words within. Clichés can

therefore be argued to be proto-idioms in the process of developing an idiomatic meaning ('Without a lawyer, the suspect was only prepared to give me the name, rank and serial number treatment'). Additionally, partial constructions (Goldberg 1995, 2003) are meaning-form pairings which have partially filled 'slots' such as *the Xer the Yer* (*the more you think about it, the less you understand, the faster the better, the bigger they are the harder they fall*, etc.) and idioms like *jog X memory* (where X could be a possessive determiner – *jog his memory* – or a genitive phrase – *jog Meg's memory, jog your friend's memory*) and so store single meanings with optional changes. Common intertextual quotes (*it is a truth universally acknowledged, hope springs eternal, a blaze of glory, to boldly go where no one has gone before*) also serve a unified function and could be usefully retrieved and analysed as single chunks (see Hüning and Schlücker 2015; Wray 2002, and chapter 7 of Jackendoff 1997, under 'If It Isn't Lexical, What Is It?' for further discussion of these and similar phenomena). All of these can have a utility in the digital humanities and avoid time-consuming manual work, but are often placed outside the MWE field. As it is, the reliance on the digital-orthographic word in corpus research often requires a significant amount of work reading concordances to identify and remove such fixed phrases and elements of (semi-) MWEs from corpus data; a process which is almost impossible to automate due to the difficulty of compiling comprehensive lists of all possible examples and the extent to which human annotators would disagree with each other over what does and does not fall into these categories.

The challenges of lexis

The biggest issue in research on lexis relates to the practice of using word frequency as a measure of a word's importance in a text. This is based on the widely accepted hypothesis that the number of times a word is present in a given text is indicative of the importance of the concept to which that word refers, at least within the text under investigation. For example, imagine two texts of the same length; in one of these, the word *custard* appears 20 times, while in the other it only occurs 5 times. Intuitively, the concept which is represented using the word *custard* would therefore seem to be more important to the text in which it is more prevalent. However, there are concerns beyond producing simple counts of word frequency; the distribution of words in written language follows a Zipfian curve, so that 80 percent of any text is composed of around 20 percent of the text's stock of words (cf. Baayen 2001: 13–17; Moreno-Sánchez et al. 2016). In this 20 percent, the upper end of the frequency list mostly comprises what may be termed 'closed-class' or 'function' words. These typically include determiners, prepositions, conjunctions and auxiliary verbs and are widely considered to contribute to the structure of a sentence and aid in the conveyance of information while carrying very little semantic content in their own right (cf. Quirk et al. 1985: 67–72). The same function words occur in a very similar frequency ranking in any large collection of English: almost always *the*, *of* and *and* in that order, followed by *a*, *in* and *to* in any order, then normally followed by *I*, *is* and *that*. Below this top fifth, the remaining text is composed of lower-frequency instances of the remaining 80 percent of words, where 'open-class' or 'content' words predominate. These are parts of speech such as nouns, verbs, adjectives and adverbs which convey most of the meaning in the text, and which are open to supplementation as new words are coined to express novel or more nuanced meanings (Quirk et al. 1985: 72). As a result of this distribution, a simple count of word frequency in most texts would return very similar results; whether analysing Dickens or Dawkins, the highest-frequency items will consist of the words given earlier in slightly varying combinations as a restricted

number of function words act as the glue which holds together content words whose supply is limited only by the creative powers of language users. These high-ranking function words are of interest for study of the general composition of the language and the way in which information is structured, but are of little use in making deductions about the focus or features of specific texts.

Rather than count the pure frequency of a word's use, therefore, it is helpful to measure whether content-carrying lexical items appear more or less frequently than would be expected in, say, a particular genre of text, a type of conversation or in a large body of text aiming to simulate general non-specific language use. To this end, many studies which draw evidence from word frequencies employ a 'reference corpus', a large body of text which is intended to provide a standard set of word usage frequencies, often for the language as a whole. In order to approach representativeness, reference corpora are preferably as large and as balanced across genres as possible. Frequently used reference corpora include the British National Corpus (BNC, version 3 2007–100 million words) and the Corpus of Contemporary American English (COCA – 450 million words), which have the advantages of being large, but are accompanied by restrictions necessitated by the manner in which they were collected. Both corpora, for example, are largely composed respectively of British and North American English usage and so are appropriate touchstones for their own varieties of English but not for English as a whole. In addition, they are not systematically enlarged with new material, some supplements aside, meaning that they drift further and further from contemporary English usage. Another problem shared by the effort to collect large corpora of current language usage is that much written material, such as novels, screenplays and print journalism are under copyright and thus are not available (or, at least, would be prohibitively expensive) for inclusion in large, widely disseminated corpora. None of these concerns invalidate these corpora, and they are excellent for particular types of research so long as they are well matched to the investigation's requirements. Even the advantage of using reference corpora containing hundreds of millions of words has been challenged recently, with the suggestion that smaller, well-designed corpora are equally useful (Scott 2009).

The challenge, then, in frequency analysis is that data in a Zipfian form is hard to compare: at the top of the distribution, where a few words account for the majority of the text, the raw frequency numbers are very high and the gap between the first and third most frequent words can be many thousands of occurrences; at the other end of the distribution, where a long 'tail' of words which occur only once (called *hapax legomena*) is found, there is no difference in occurrence number. What this means in practice is that, in a corpus of a few million words, an increase of 1000 occurrences will make no difference in the frequency ranking at all for items at the top of the curve but will result in an enormous jump in ranking for those at the bottom; in contrast, a change of, say, 20 in frequency rank will require many thousands more occurrences at the top of the curve but could be achieved by two or three extra occurrences at the base of the curve (for more, see Sorell 2015). Statistical measures must be used to therefore account for the nature of the Zipfian layout in order to find words in a text which occur proportionally more frequently than they do in the reference corpus. Often, this significant deviation from expectations is indicative of the focus, or 'about-ness', of that text (see Bondi and Scott 2010).

Of great importance to the study of lexis in use is its surrounding context, as is demonstrated by the long tradition of manual concordance production for culturally important texts such as biblical and patristic writing, classical literature or the works of Shakespeare. Concordances, both in the digital humanities and in physical format, generally take a 'key word in context' (KWIC) form, in which a short excerpt of the text surrounding and centred

on the desired word is presented. The vocabulary which surrounds a word in use is indicative of the associations held with that word, either by language users in general or by the author or authors of the texts under study. Collocation statistics – measures of the degree to which one lexical item can be expected to appear in the vicinity of another – can offer important insight into the connotations which surround words; as John R. Firth once wrote, ‘the complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously’ (1935: 37) – or, in a more famous quotation, ‘you shall know a word by the company it keeps’ (1962 [1957]: 11). A related notion is *colligation*, where words are described in terms of their co-occurrence with grammatical patterning (verbs of perception, for example, colligate with an object followed by a verb in either the bare infinitive or -ing form: *we saw them running, I heard the clock strike, they witnessed the murder happening*). Critical to the development of what are today thought of as colligation patterns was the work of John Sinclair, who argued that lexis and grammar are not separable features of language and that a word and the lexical and grammatical structures in which it is embedded themselves form a unit of meaning that is more abstract than linguistics had previously generally recognised (Sinclair 1991: 81 ff.).

Through an examination of collocation patterns, a picture of the linguistic environment of lexical items can be constructed, and the resultant associations can form the *semantic prosody* of the item; this is ‘a form of meaning which is established through the proximity of a consistent series of collocates’ (Louw 2000: 57). This prosody can show the associations of a term which a language user holds below the level of conscious thought and is therefore often argued to express an unconscious attitude or evaluation of a speaker, or to demonstrate irony, insincerity or humour (Louw 1993). Where a word has, for example, largely negative connotations, this can be evidenced in its frequent collocates; the phrasal verb *set in* is a frequently used example here (cf. Louw 1993; Sinclair 1991: 70–79), where the literal meaning is simply ‘develop’, but analysis has shown that *set in* is almost always found with negative collocates, such as *rot, decay, despair, rigor mortis, prejudice* and so on. The argument then goes that a statement of the meaning of *set in* which does not include an awareness of the semantic prosody and its negative focus would result in an insufficient definition. Studies such as that of Gabrielatos and Baker (2008) have examined such issues as the representation of refugees and asylum seekers in press coverage, in this instance establishing that terms such as ‘refugee’ have collocation patterns in British tabloid and conservative broadsheet newspapers which reinforce negative stereotypes of refugees, asylum seekers, immigrants, and migrants; for example, such individuals are portrayed as arriving in large number (the authors identify collocates including *flooding, pouring* and *streaming*), causing economic problems (e.g. *benefits, claiming*) and are treated as lawbreakers (e.g. *caught, detained* and *smuggled*; for all examples see Gabrielatos and Baker 2008: 21) This is evidence of the semantic associations of words in specific, press-related contexts, whereas studies employing more general, balanced corpora produce more overarching data about a word’s use.

Data issues

In lexis-based research, complicating factors are introduced by the nature of the data. The issue of polysemy is of prime importance, as it is regularly the case that a researcher is really only interested in a single meaning of a polysemous lexical item. We are faced frequently with the need to distinguish individual senses, which can be complicated further by the distinction between polysemy and homonymy. Polysemy – the existence of multiple distinct

but semantically related senses for the same word spelling – can be important for the study of word formation through association and metaphorical extension, whereas homonymy – the representation of unrelated ideas using the same word spelling – confuses this picture (Lyons 1977: 550–552). Often the use of appropriate collocates can help (for example, in a historical corpus such as the Corpus of Historical American English, examining concordance lines for *gay* and comparing those for the collocate *lesbian* with those for the collocate *happy* gives a very rough means of distinguishing the newer and older senses). In addition, methodologies for automatically distinguishing polysemes and homonyms are improving, and are discussed briefly later.

A special category of the issue of polysemy is that of metaphor. Metaphorical senses, including very common expressions like TIME IS SPACE (*That's all behind us now, Next week and the week following it*), HAVING CONTROL IS BEING HIGHER UP (*I am on top of the situation, He's at the height of his power, He is under my control, He is my social inferior*) and so on all add further semi-polysemous senses to any lexical search, as recent research has shown that between 8 percent and 18 percent of English discourse is metaphorical (Steen et al. 2010: 765; all the preceding examples are from Lakoff & Johnson 1980, and for a more detailed discussion of metaphorical lexis in English within DH, see Alexander and Bramwell 2014).

Parallel to the challenges of words with the same spelling and different meanings are the presence of multiple spellings intended to represent the same lexical item. Historically, the lack of standardised spellings preceding the late modern period is particularly problematic, so that in Early Modern English, for example, the word *would* may be found with the spellings *wolde*, *woolde*, *wuld*, *wud*, *wald*, *vvould* and *vvold* amongst others (Archer et al. 2015: 9). Digital methods for associating variant spellings such as these are of great importance to the future of historical philology; textual analysis software is usually more effective if variant spellings can be combined under the same standardised token, especially where frequency of word usage is a primary concern. It is worth noting, however, that some researchers have seen value in maintaining the distinction between historical spelling variants, not only for study of dialectal difference but also for its potential effect on representation of polysemy. Variant spelling is not purely a concern for historical research; users of corpora based on Internet data, for example, require the ability to recognise typographical mistakes and unconventional spellings as representing the same word as the standardised spelling. Indeed, spelling differences between different varieties of English – such as the well-known distinctions between U.S. English and British English (e.g. *color* versus *colour*), or the Oxford and American use of the *-ize* spelling versus the alternative *-ise* spelling – may require standardisation for effective automatic analysis. Spelling normalisation programmes such as VARD (VARIANT Detector) developed at Lancaster University's UCREL unit may be used to homogenise the spelling of historical texts. Such systems tend to be underpinned by a set of pre-identified rules (e.g. 'change word-final – *ck* to – *k*, thus *physick* to *physic*') and known word-specific variants (e.g. 'change *wolde* to *would*'), although VARD allows the user to 'train' the system on a section of data before an automated process is deployed (Baron and Rayson 2008).

Distinguishing components of data, such as polysemy or spelling variants, may either rely on or add to the layer of metadata (that is, data about the data) which is associated with the text in question. General metadata, potentially including such information as place of production and author's name, is of special importance for filtering data to meet controlled variables in research. The type of metadata recorded varies between text collections depending on the needs or projected needs of their users, although an overall

scheme for structuring any gathered metadata in text collections is curated by the Text Encoding Initiative (TEI). Metadata can, however, be problematic, especially where time or budgetary constraints have limited precision: for example, Early English Books Online (EEBO) consists of digitised microfilm images of printed material produced c.1475–1790 CE; however, largely due to the nature of the source material, not all of the texts in EEBO have full data for author and place of publication, and sometimes this information has been copied verbatim from the texts so that *London* may appear as *Londini* or *Impressa Londini* where a Latin colophon has been employed. Also usually present for linguistic research is metadata relevant to individual words. This may include the word's part of speech (POS) and lemmatised form (see the earlier section), based on automatic morphological and syntactical analysis, with some systems additionally employing spelling normalisation procedures.

Current contributions and research

While there are many challenges for digital humanities scholars in working with lexical data, as outlined earlier, in this section we overview the major sources of information about lexis which help us overcome them. We take here the perspective of degrees of 'curation' of data, and so begin with the major carefully created sets of data about lexis – including dictionaries and thesauri – and then discuss the major tools which can be applied to the study of lexis, focusing on semantic tagging software, lemmatisers and spelling normalisers.

Research can begin from any one of these resources. Current DH tools and resources allow a question to be asked on the basis of interesting features arising in lexicographical data (such as a dictionary or thesaurus) with the researcher then proceeding to corpora to test their hypotheses there; conversely, a question may arise from analysis of corpus data which then requires information (such as pronunciation, etymology or spelling variation) from lexicographical resources in the process of formulating an explanation or hypothesis to explain the data. This distinction is often formalised as a tension between *corpus-based* research (i.e. testing hypotheses on corpus data) and *corpus-driven* research (allowing patterns to emerge from data which are then investigated; for further discussion see Tognini-Bonelli 2001: 65–100). Neither is to be preferred over the other, and in practice, research often cycles through these as alternating phases within the same project.

Resources

The current major contributions to and resources for lexical research can be divided along lines which relate to the nature of the resource and the degree to which it has been curated. Resources which have involved high degrees of editorial intervention include the major dictionaries and thesauri and contain carefully collected and maintained data about lexemes themselves; as the degree of structure and curation decreases, we travel along a cline from information to data, increasing the 'raw'-ness of the lexical data being used. The rawest data which can be used as a resource are an unstructured web-as-corpus dataset; from there corpora increase in detail and complexity of their structure and metadata until we reach highly curated and already comprehensively researched lexicographical resources. As DH and corpus linguistics research has progressed, digital corpus-based datasets have increasingly fed back into the dictionaries and thesauri which constitute the highly structured end of this cline, lending a symbiotic aspect to their relationship.

Major lexicographical resources include:

- The *Oxford English Dictionary* (OED; www.oed.com). The pre-eminent English dictionary created on historical principles. The second edition and its supplements were published in 1989 and form the core of the currently available online edition. Thorough revision of the dictionary is underway with regular updates bringing entries of the online version to a third edition status. The OED data include dates of attestation for its entries with accompanying source material quotations, lists of variant spelling forms and etymological data. The current revision includes piloting frequency data from historical corpora, which should give an indication of whether there is evidence of a word (but at present not a sense) being rarely or heavily used.
- Online OED-style (scholarly, multi-volume, organised on similar principles and featuring citation evidence) dictionaries exist for particular varieties or periods of the language. These include:
 - The *Dictionary of the Scots Language* (DSL; www.dsl.ac.uk). A historical dictionary, drawing on the *Dictionary of the Older Scots Tongue* (DOST) and the *Scottish National Dictionary* (SND) to provide an electronic resource of more than 80,000 entries.
 - The *English Dialect Dictionary* (EDD; <http://eddonline-proj.uibk.ac.at/edd/>). An online edition of Joseph Wright's six volume 1898–1905 dictionary of words found in English dialects over 200 years, based on the publications of the English Dialect Society.
 - The *Dictionary of Old English* (DOE; www.doe.utoronto.ca/). This is in progress and covers the period 600–1150. Already available is A–I.
 - The *Middle English Dictionary* (MED; <https://quod.lib.umich.edu/m/med/>). The sister publication of the DOE, covering the period 1100–1500, completed in 2001 and made available online from 2007.
 - *A Dictionary of American English on Historical Principles* (DAE; not available as a database but available as page scans through the Hathi Trust web pages). Compiled in the 1930s, DAE covers English in America from the earliest settlements to the twentieth century.
 - The *Dictionary of American Regional English* (DARE; www.daredictionary.com/). Published between 1985 and 2013, DARE does not document standard American English, but instead records dialectal words and includes dialect maps.
 - The *Australian National Dictionary* (AND; <http://australiannationaldictionary.com.au/>; only the first edition is online at present), first edition 1988 and second edition 2016. A historical dictionary of Australian English, covering Australian words not found in general English dictionaries.
- Many other online dictionaries are available, such as *Wordnik* (<http://wordnik.com/>), an aggregator of various lexical resources, including out-of-print dictionaries and examples gathered from social media and elsewhere. Many of the other online dictionaries are of varying quality, however, and should be used with care.
- The *Historical Thesaurus of English* (HTE; <http://ht.ac.uk/>) organises the lexis of English into a detailed semantic hierarchy beginning with three major divisions – *The World*, *The Mind* and *Society* – each subdivided into categories such as *Health and Disease*, *Attention and Judgement* and *Occupation and Work*, themselves divided to a maximum of seven levels. The data of the HTE is largely drawn from the second edition of the OED

and its supplements, with the addition of Old English material gathered for *A Thesaurus of Old English* (<http://oldenglishthesaurus.arts.gla.ac.uk/>), and work is currently underway on a second edition. The semantic categorisation of the HT enables historical semantic research which would otherwise have been prohibitively time-consuming or, indeed, impossible.

- WordNet (<https://wordnet.princeton.edu/>) is a resource which links together words of similar meaning into groups known as *synsets*, while also noting some limited semantic relationships such as hyponymy (e.g. that a *ship* is a type of *vehicle*) between the synsets. One of WordNet's advantages is that it is freely downloadable, which has led to its wide use in computational studies, although its quality and comprehensiveness are not of the same order as the other resources described here.

Alongside lexicographical resources provided by dictionaries and thesauri, large textual corpora are of great importance to the study of language in use. Although many corpora are constructed as small highly curated sources for study of very specific data, corpus resources are increasing in size at a rapid rate, with next-generation corpora often outpacing the previous decade's resources by an order of magnitude (cf. Sinclair 1991: 1; Davies 2015: 12–13). Corpus resources are essential for giving data on lexis in use. Some major corpora include the *BNC*, the *COCA*, the *Corpus of Historical American English*, the *International Corpus of English* and more (see Anderson and Corbett 2017). A key historical corpus is *EEBO*, already mentioned earlier, which gathers a vast amount of material printed in English or in English-speaking countries in the early modern period (roughly 1475–1700 CE). Around 125,000 sources are represented in *EEBO* and are available through subscription-based resources as scans of microfilm page images. Hand-typed Extensible Markup Language (XML) transcriptions of c. 40,000 items have been produced by the Text Creation Partnership (TCP) consortium, 25,000 of which have been made publicly available with more to follow. A semantically tagged *EEBO*, alongside the 1.6-billion-word semantic *Hansard* corpus (including 200 years of UK Parliamentary debates) was released in 2017.

Tools

An array of tools employs the lexical data compiled and curated in the previously mentioned resources. The practice of digital textual analysis employs these tools to enrich a given text with metadata and annotations, enhancing the abilities of researchers to study grammatical, semantic and discursal patterns in lexical data. The technology for many of these applications is still in its infancy but opens exciting new possibilities for research on lexis and its use.

- The UCREL Semantic Analysis System (USAS) is a text-tagging pipeline developed at Lancaster University's University Centre for Computer Corpus Research (UCREL). It normalises the spelling of any text with which it is provided, using the UCREL VARD tool and then assigns POS and lemma tags to words in the text along with a semantic label based on classification of words into broad thesaurus-style categories, thus enabling analysis of key semantic fields and semantic collocation on top of lexical collocation.
- The SAMUELS tagger builds on the USAS tagger, using the HTE as its source of semantic classification. The highly fine-grained nature of the HTE classification system, alongside its word-sense dating information, allows more detailed semantic tagging which considers the senses available during the period of the text's production.

This is the tool used to develop the semantic EEBO and Hansard corpora mentioned earlier, and is described in more detail in Alexander et al. 2015.

- The *Mapping Metaphor* project is, like SAMUELS, a resource developed from *Historical Thesaurus* data. Through manual comparison of HTE categories, the project has mapped the semantic categories which have acted as source and target domains for metaphorical borrowing of lexis resulting in new sense creation. Examples of its use can be seen in Anderson et al. 2016.

Connecting resources and tools

As an in-progress example, work on linking new processes with these tools is the ‘Linguistic DNA of Modern Thought’ (LDNA) project, a collaborative venture between the universities of Sheffield, Glasgow and Sussex funded by the UK AHRC. The purpose of LDNA is to use automatic text processing to reveal groups of words which co-occur frequently enough to suggest they form a coherent ‘discursive concept’ (Fitzmaurice et al. 2017). This differs to some extent from more traditional semantic analysis, which treats concepts as units of meaning which can be expressed by a single word or by a group of synonyms. As a result, where the more traditional approach would see *car*, *motor car* and *automobile* as lexical representations of a concept, the approach taken by LDNA might expect to find *car*, *exhaust*, *driver*, *pollution*, *highway* (and perhaps further items) as a group occurring with enough frequency in text to suggest the existence of a concept which embraces all of this lexis. LDNA is using the EEBO-TCP corpus (see earlier), which is not only a large textual resource but also represents a period in history in which wide-ranging societal and intellectual development might be expected to have an impact on the concepts expressed in writing. The EEBO-TCP XML text is processed to normalise historical spelling variation, lemmatise words and tag them with their POS. Once this is done, the next stage of processing is to calculate co-occurrence statistics between each word and those in a roughly paragraph-length window around it. The statistical results are then used to group together words which are strongly associated with each other. These are akin to collocates in similar corpus linguistic research, although they are acquired from larger spans of text and in larger groups than are typically employed by such work, with groupings built outwards from a seed node to word pairs, trios and eventually quads using adaptations of the pointwise mutual information (PMI) significance measure developed for the project. The aim is for the word groupings which emerge from processing to tell us more about not only the manner in which word association works but also allow us to trace conceptual development in early modern printed text. It is this representation of societal preoccupations which forms the ‘linguistic DNA of modern thought’ of the project’s title. The results of the project are freely available on the project’s website (www.linguisticdna.org) and constitute the results of the word grouping process, visualised in such a way as to encourage exploration according to the individual interests of users.

Sample analysis

In this section, we briefly demonstrate the main ‘directions’ of research techniques in this area (from lexis to corpora, from corpora to lexis or both through the perspective of a connected semantic field) by a study of lexis in a category of the *Historical Thesaurus*: the nouns in ‘03.12.15 Money’. This category has been chosen because it is highly culturally significant to present-day societies, as witnessed in the English-speaking world by the prevalence

of financial matters in print and broadcast journalism and the existence of (at least notionally) finance-focused international publications such as the *Financial Times*, the *Wall Street Journal* and the *Economist*. In addition, the category ‘03.12.15 Money’ offers over 100 near-synonymous nouns in a lexicon which demonstrates the prevalence of metaphor, slang, onomatopoeia and euphemism in vocabulary development. Money and its pursuit appears, in part, to be an especially productive category because of its ethical ambivalence; it can be said to make the world go round or viewed as the root of all evil, and has been key to human society since around 1200 BCE, when we began to move from direct barter of goods and services to abstract means of representing value through tokens (Ferguson 2008). We attempt to demonstrate here the category’s variation in terms of its meaning, evolution, internal structure, metaphorical extensions to other types of lexis and evidence of use.

The importance of money is reflected in English by the large number of words which have been used for the concept from Old English to the present day. Category ‘03.12.15 Money’ in the HTE (see Figure 10.1) provides a list of the concept’s realisations for the language’s millennium-and-a-half history. As a phenomenon arising from interaction rather than part of either the physical or mental worlds, the category ‘Money’ is found in the third major division of the HTE, glossed as ‘Society’ and is then a daughter of the second-tier category ‘3.12 Trade and Finance’. Without including subcategories, there are a total of 108

01 The world
02 The mind
03 Society

03.01 Society and the community
03.02 Inhabiting and dwelling
03.03 Armed hostility
03.04 Authority
03.05 Law
03.06 Morality
03.07 Education
03.08 Faith
03.09 Communication
03.10 Travel and travelling
03.11 Occupation and work
03.12 Trade and finance

03.12.01 Trading organization
03.12.02 Trading conditions
03.12.03 Bargaining
03.12.04 Barter
03.12.05 Brokerage
03.12.06 Illegal/immoral trading
03.12.07 Other trading methods
03.12.08 Importing/exportation

03.12.15 n. MONEY

PARTS OF SPEECH: n. Money adj. Pertaining to money
phr. Money

There are 108 words at this level, sorted by first attested date:

mynet OE
scat < sceatt OE-C1250
fee < feoh OE-1677 + 1823- slang
pence < peningas OE-
silver < seolfor OE-
spense a1225-C1460
spending a1290-1650
money c1290-
worth a1300
muck a1300-1710
cattle c1330-1483
penny c1330-

Figure 10.1 Section of the Historical Thesaurus category ‘03.12.15 Money’ from the Thesaurus’ website

nouns in the ‘Money’ category, 5 of which originate in Old English – *mynet*, *scat* < *sceatt*, *fee* < *feoh*, *pence* < *peningas*, and *silver* < *seolfor* (a chevron separates an OED headword from its Old English ancestor word). Three nouns (*fee*, *pence* and *silver*) are noted as still active in present-day English (PDE). Each century (with the exception of the fifteenth, although Middle English dating is notoriously challenging) add items to the list, and *money* itself is coined in c1290. Of the 108 items, 65 are recorded as still in use, and of these 65, all but 13 are marked as being slang or colloquial terms.

The high prevalence of slang terms can be at least partially accounted for by the unusual status of money as both important to survival and simultaneously an almost taboo subject. The tendency to avoid direct reference to it in conversation has encouraged a variety of euphemisms to emerge which allow English speakers to talk about money in an evasive or jocular manner, distancing themselves from or making light of the subject’s negative cultural associations. A number of metaphors imply a corruptive or polluting influence, such as *muck* (a1300–1710), *trash* (a1592–1809) and *the filthy* (1877 + 1931–present); some note its importance, such as *the necessary* (1772–present) and *the needful* (1774); some are clearly metonymies derived from the material used to make coins, such as *silver* < *seolfor* (OE-present), *brass* (1597/8–present, and *pewter* (1829–1888); while still others are more opaque, such as *gingerbread* (a1700–present, likely from the way gingerbread biscuits were often gilded historically), *mint-sauce* (1828–present, a pun on *mint* meaning money, usually now only found as the term for the process of making coins), *soap* (1860–present, potentially a reference to money’s ability to lubricate otherwise insoluble situations) and *sugar* (1869–present, whose origin is obscure but may conflate the commodity of the formerly lucrative sugar trade with the wealth in which it resulted, relying on the overlapping properties of both to facilitate pleasure).

Although this batch of metaphors is selective, it suggests a relationship between the domains of MONEY and FOOD, perhaps as both are finite commodities essential to life whose acquisition nonetheless requires a significant expenditure of effort. It is possible to measure this hypothesis against the data produced by the *Mapping Metaphor* project, which has systematically identified instances of lexical metaphors which connect semantic fields in the *Historical Thesaurus*. Using this project’s website, a search for the term *money* in category headings returns category ‘3L02 Money’, and amongst those categories which are linked to ‘Money’ is, indeed, ‘Food and eating’ (1G01, see Figure 10.2). The data available here identify a strong link between the two, with ‘Money’ drawing from ‘Food and eating’ as a source domain since the first half of the sixteenth century. Amongst the sample items for the ‘Money’ and ‘Food’ link are *from hand to mouth*, *dough* and *uncooked*. There are, however, many other categories linked to ‘Money’. ‘Money’ has, for example, drawn on ‘1J14 Liquid’ for items such as *liquidate*, *melt* and *unfreeze*, ‘2B06 Unimportance’ has taken ‘Money’ as its source for terms such as *halfpenny*, *two pennyworth* and *at a pin’s fee*, and ‘1I04 Weariness’ has a two-way connection, including the items *spend*, *exhaust* and *overtax*.

Returning to lexicographical resources, it is also worth selecting a single term and drilling down into the data available. Of the terms which have been present for several centuries, *cash* (1651–present) is perhaps the most readily recognisable. It occurs within the *Historical Thesaurus* category ‘03.12.15 Money’ in the main body of the category as a term for money, but also in the sub-category ‘03.12.15|03 ready money/cash’. Subcategories are used to house senses which have specific, fine-grained meaning relevant to the category. Inclusion in a sub-category does not, however, imply lack of importance, and this sub-category sense of *cash* is in fact the more prevalent usage of the term.

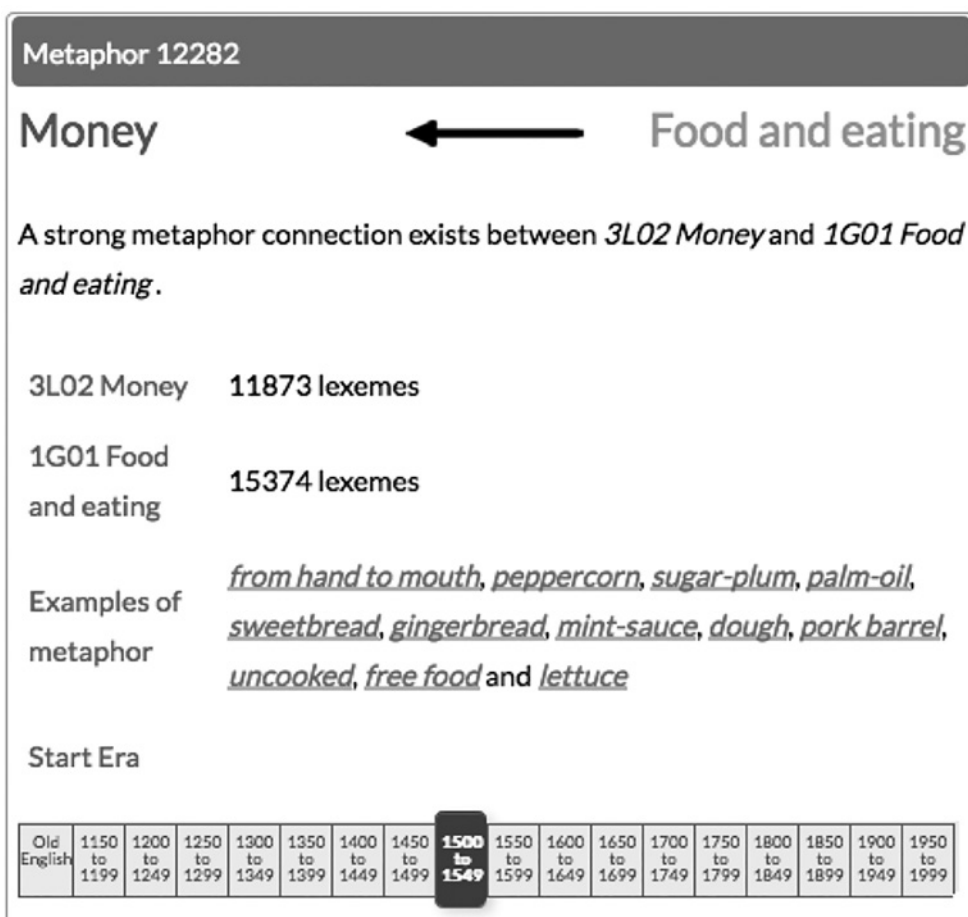


Figure 10.2 Metaphor 'card' showing links between 'Money' and 'Food and eating' from the Mapping Metaphor project website

Source: www.glasgow.ac.uk/metaphor

Cross-referenced in the OED, there are two entries for *cash* – 'cash, *n.1*' is the appropriate entry for the sense which is near-synonymous with *money*, although 'cash, *n.2*' is also monetary, where the word denotes coins in particular financial systems of the East Indies, southern India and China. The etymological data provided by the OED makes clear that the *money* sense of *cash* is the result of metonymic extension; there are two possible sources from which the word may have been borrowed, French *casse* and Italian *cassa* (ultimately derived from Latin *capsa*, coffer) both of which denote a chest or similar receptacle used for storage of money rather than the money itself. The editors indeed remark that 'the sense "money" also occurs notably early, seeing that it is not in the other languages'. The now obsolete 'money box' sense of *cash* is the earliest recorded by the OED, with the first illustrative quotation taken from an account of Sir Francis Drake's voyage to the Americas by Thomas Maynarde (c1595). The sense of 'Money; in the form of coin, ready money' – equating to the HTE's sub-category 'ready money' sense – is the second of the dictionary's three main senses of *cash*.

There is also a third listing; an adjectival use of the noun in relation to ‘(a) commodities purchasable for cash, (b) tradesmen or commercial houses doing business for ready money only’. The meaning present in the main HTE category ‘03.12.15 Money’ is to be found in OED sub-sense 2.d, which constitutes ‘the regular term for “money” in Book-keeping’.

If one’s interests lie away from etymology, it is possible to better understand the semantics of ‘cash’ through evidence in context from corpora. For example, EEBO may be accessed through the CQPweb resource provided by Lancaster University (<https://cqpweb.lancs.ac.uk>), a resource which returns KWIC concordance lines with the option to limit searches by text-level metadata such as publication date and word-level metadata such as part of speech. In order to find noun forms only (as opposed to, say, verb instances such as ‘cash a cheque’), the part of speech tag ‘NN1’ – singular noun – can be employed as part of the search term: ‘cash_NN1’. This returns 1,437 matches across 477 texts and a usage frequency of 1.20 instances per million words. This identifies ‘cash’ as a much less commonly employed term than ‘money_NN1’, at least in the period covered by EEBO (approximately 1475–1700) which has a frequency of 156.43 instances per million words. The breakdown of distribution across time clearly supports the dating information provided for the noun *cash* in the OED and the HTE, which date its earliest uses to 1596; the CQPweb EEBO returns only nine instances across six texts in the period prior to 1600, with a rapid increase to 1349 instances in 606 texts for the seventeenth century.

Using the tools available for analysis via CQPweb, the most frequent collocates of *cash* are found to be *ready* and *running*. The tool also returns a log-likelihood (LL) score for these collocations, a statistical measure of confidence that the results are not occurring by chance alone; *ready* and *running* score 768.838 and 645.674, respectively, comfortably above the threshold of 16 or so which indicates significant results (Rayson et al. 2004). The co-occurrence of *ready* and *cash* (in fact, 92.7 percent of these instances consist of the phrase ‘ready cash’ with words in that order) is indicative of the use of *cash* to denote money which can be accessed instantly for use, with *running* (here 100 percent of instances are in the phrase ‘running cash’) acting as a more corporate reinforcement. Beyond *ready* and *running*, top of the list of *cash* collocates in EEBO are *debtor* (LL 373.475) and *creditor* (LL 232.772), which reflect the propensity for the word to appear in the context of money lending. CQPweb allows the user to view the instances it has retrieved in context by producing concordance lines which show a number of words to either side of the search word. Although sifting through these manually can be time-consuming, it may yield results which suggest avenues for further investigation.

It is also possible to use CQPweb to broaden our scope again from *cash* to *money* lexis more generally. The CQPweb instance of EEBO-TCP allows semantic searching using the USAS semantic tagset, where words which are identified by the system as relating to ‘Money generally’ are labelled ‘I1’, with subdivisions for ‘Money: affluence’ (I1.1), ‘Money: Debts’ (I1.2), and ‘Money: Price’ (I1.3). The same is also possible using a more fine-grained system with texts tagged by the Historical Thesaurus Semantic Tagger (HTST) developed by the SAMUELS project. The HTST owes much of its architecture to the USAS tagger but employs the much more detailed and rigorous architecture of the *Historical Thesaurus of English*, as well as its sense-dating information. The two centuries of the *Hansard Corpus* can currently be searched for SAMUELS semantic tags on the English-corpora.org website (www.english-corpora.org/hansard). The interface has a built-in option to search for semantic categories and offers the user both a list of broad semantic categories to choose from and the ability to search a larger set of category titles for more specific information. In this structure, the category ‘Money’ is given the identifying tag-code ‘BJ : 01 : m’, which is part of the *Historical Thesaurus* hierarchy structure. This code can, in turn, be used as a search term

SEE CONTEXT: CLICK ON WORD (ALL SECTIONS), NUMBER (ONE SECTION), OR [CONTEXT] (SELECT) [HELP..

		CONTEXT	ALL	1800	1810	1820	1830	1840	1850	1860	1870	1880	1890
1	☐	MONEY	820080	2080	2516	5585	9639	11809	13579	16677	16275	29817	319
2	☐	TREASURY	271484	717	424	759	2181	2881	4942	6332	7568	21941	269
3	☐	SUM	186996	1187	2312	3073	6186	6176	8698	9625	8144	14383	125
4	☐	INFLATION	71217						5		13	8	8
5	☐	SUMS	68973	470	494	743	1321	1618	1799	2165	1774	2974	261
6	☐	THOUSANDS	67898	76	171	244	738	1126	867	1001	1067	1965	165
7	☐	CURRENCY	45427	108	1094	1994	1508	1058	578	304	403	330	99

Figure 10.3 Top search results for semantic category 'BJ : 01 : m – Money' in the semantically tagged Hansard Corpus

which will return results for all of the tokens in the corpus which the SAMUELS system has tagged as belonging in this category. Figure 10.3 displays the results of such a search, giving the lexeme returned in column 3 (such as *money*, *treasury* and *sum*) ranked by the frequency of their occurrence in column 4. Remaining columns are headed by the decade to which texts are dated and give figures for each lexeme's occurrence in the appropriate time period.

The ability to search for semantic tags rather than individual words is one of the important ways in which current digital humanities methods are beginning to allow more meaningful searching, allowing searches for meaning rather than the more time-consuming use of keyword search terms (cf. earlier in this chapter). A search for tag 'BJ : 01 : m' in the Hansard corpus therefore returns results for all of the lexical items related to 'money' rapidly and efficiently – a result which previously would have required the creation of a list of search terms followed by filtering of the results to remove irrelevant meanings of polysemous terms. While this technology is by no means perfect as yet – the current accuracy of the HTST is 79.85 percent when averaged over test files of different text types and composition dates (Piao et al. 2017: 126–127) – it is taking dramatic steps forward and is likely to be a major component of future data retrieval, with some tools allowing users the option to manually correct inaccurate tagging in the interim.

It is therefore remarkably quick and easy to gather information from a variety of academic resources which give insight to the history, etymology, semantics and usage of lexical items in the semantic field of 'Money'. Although little more than a pen-sketch outline can be provided here, the resources currently available would allow any of these points to be pursued to much greater depth. Future resources should allow us to take another step in tracing the development of the semantic field and its related concepts and lexis, and the future of such lexical research should open up yet further in response to such tools and data.

Future directions

Current research on lexis has a strong interest in investigating the manner in which context affects the usage and understanding of words. Semantic research is likely to continue on this track for the near future, using statistically informed methods to map networks of association between lexical items. Such work rests on the Firthian belief that word meanings are

in large part a product of the circumstances in which they are used. This is closely related to the idea of semantic prosody, which seeks word connotations through patterns of lexical associations. Techniques such as cluster analysis and network analysis are at the forefront of this type of research and have pushed forward not only the knowledge of word association patterns but methods by which these patterns may be visualised in ways more readily interpretable to human eyes than tables of figures. An example of new work on this is given earlier with reference to the *Linguistic DNA* project.

Many other projects exist working on lexis, though – in fact, very few current DH projects in English do not use lexis as a key component – and the rest of this volume repeatedly illustrates this. Working towards the future, a clear issue which arises when looking at these projects is that the ever-present problems of polysemy, homonymy and precision mean it is very difficult to link such resources together. One word in context cannot easily be tied to a meaning or other data about that word: a search for the word form *bow* in the HTE, for example, returns 72 possible meanings, and so KWIC results for *bow* in a corpus cannot be linked to further information such as etymology (depending on the sense, it could be from Old English or Old Norse), pronunciation (again, could be bæw or bau), semantic field and so on. This requires significant manual work to resolve. The HTE and the OED, for example, are very closely linked, and this linkage allows the connection of word senses in corpora tagged using a tagger such as the HTST, but this is a rare instance of feasible interconnections between resources. There is also, as yet, no unique identifier for a word sense: the *Thesaurus* uses OED sense ID fields to link its data, and OED senses are potentially the best resource for precisely identifying a word sense within an etymological, phonetic and semantic context – but these are not easily found or reused. This, it should be said, is no criticism of the OED, but rather an indication of the gaps in the field of lexis more generally when it comes to taking what we already know and connecting it with more and more resources to gather greater and deeper insights, in the best traditions of the DH.

Further reading

- 1 Taylor, J. (ed.) (2015). *The Oxford handbook of the word*. Oxford: Oxford University Press.
This book gives dozens of perspectives on words, many supported with digital resources, and can inspire much thought for research and further study. Topics range from *which words do you need*, *do words exist* and *how many words are there* to *words as carriers of cultural meaning*, *word puzzles*, *the word as a universal category* and *how infants find words*.
- 2 Kay, C. and Allan, K. (2015). *English historical semantics*. Edinburgh: Edinburgh University Press.
This textbook is an outstanding summary of the study of meaning in English, written by leading practitioners. Chapters 4 and 6 give excellent information on how to carry out studies in historical meaning and evolution using digital methods, particularly the OED and *Historical Thesaurus*.
- 3 Allan, K. and Robinson, J. (eds.) (2011). *Current methods in historical semantics*. Berlin: Mouton de Gruyter.
The chapters in this book focus on data-driven investigations into meaning and lexis, including commentaries on data and sources, corpus-based methods and theoretical approaches.
- 4 Bird, S., Klein, E. and Loper, E. (2009). *Natural language processing with python – analyzing text with the natural language toolkit*. Sebastopol, CA: O'Reilly. Revised edition available as open access from www.nltk.org/book/.
Still the best way to get started with using a programming language – in this case, the free and widely used Python language and the also-free NLTK package – to work on lexis and texts. Also excellent on managing data, computational methods, writing programmes and fundamental language processing techniques.

References

- Alexander, M. and Bramwell, E. (2014). Mapping metaphors of wealth and want: A digital approach. *Studies in the Digital Humanities* 1: 2–19.
- Alexander, M., Dallachy, F., Piao, S., Baron, A. and Rayson, P. (2015). Metaphor, popular science, and semantic tagging: Distant reading with the *Historical Thesaurus of English*. *Digital Scholarship in the Humanities* 30(s1): i16–i27.
- Alexander, M. and Struan, A. (2013). ‘In Countries So Uncivilized as Those?’: The language of incivility and the British experience of the world. In M. Farr and X. Guégan (eds.), *The British abroad since the eighteenth century* (Vol. 2). London: Palgrave Macmillan.
- Anderson, W., Bramwell, E. and Hough, C. (2016). *Mapping English metaphor through time*. Oxford: Oxford University Press.
- Anderson, W. and Corbett, J. (2017). *Exploring English with online corpora* (2nd ed.). Houndmills: Palgrave Macmillan.
- Archer, D., Kytö, M., Baron, A. and Rayson, P. (2015). Guidelines for normalising early modern English corpora: Decisions and justifications. *ICAME Journal* 39: 5–24.
- Baayen, R.H. (2001). *Word frequency distributions* (Vol. 1). Dordrecht: Kluwer Academic Publishers.
- Baker, H. and McEnery, A. (2017). *Corpus linguistics and 17th-century prostitution: Computational linguistics and history*. London: Bloomsbury.
- Baron, A. and Rayson, P. (2008). *VARD 2: A tool for dealing with spelling variation in historical corpora*. In *Proceedings of the postgraduate conference in corpus linguistics*. Birmingham: Aston University, 22 May. Available at: <http://eprints.lancs.ac.uk/41666/1/BaronRaysonAston2008.pdf>.
- Bondi, M. and Scott, M. (eds.) (2010). *Keyness in texts*. Amsterdam: John Benjamins.
- Champion, E.M. (2017). Digital humanities is text heavy, visualization light, and simulation poor. *Digital Scholarship in the Humanities* 32(s1): i25–i32.
- Davies, M. (2015). Corpora: An introduction. In D. Biber and R. Reppen (eds.), *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press, pp. 11–31.
- Durkin, P. (2009). *The Oxford guide to etymology*. Oxford: Oxford University Press.
- Ferguson, N. (2008). *The ascent of money: A financial history of the world*. London: Penguin.
- Firth, J.R. (1935). The technique of semantics. *Transactions of the Philological Society* 34(1): 36–73.
- Firth, J.R. (1962 [1957]). A synopsis of linguistic theory 1930–1955. In *Studies in linguistic analysis* [Special volume of the Philological Society]. Oxford: Blackwell, pp. 1–32.
- Fitzmaurice, S., Robinson, J.A., Alexander, M., Hine, I., Mehl, S. and Dallachy, F. (2017). Linguistic DNA: Investigating conceptual change in early modern English discourse. *Studia Neophilologica* 89(s1): 21–38.
- Gabrielatos, C. and Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996–2005. *Journal of English Linguistics* 36(1): 5–38.
- Gibbs, R.W. (1985). On the process of understanding idioms. *Journal of Psycholinguistic Research* 14(5): 465–472.
- Goldberg, A.E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago, IL: Chicago University Press.
- Goldberg, A.E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7(5): 219–224.
- Hockey, S. (2004). The history of humanities computing. In S. Schreibman, R. Siemens and J. Unsworth (eds.), *A companion to digital humanities*. Oxford: Blackwell, pp. 1–19.
- Hüning, M. and Schlücker, B. (2015). Multi-word expressions. In P.O. Müller, I. Ohnheiser, S. Olsen and F. Rainer (eds.), *An international handbook of the languages of Europe, volume 1*. Berlin: De Gruyter, pp. 450–467.
- Jackendoff, R. (1997). *The architecture of the language faculty*. Cambridge, MA: MIT Press.
- Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: University of Chicago Press.

- Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis and E. Tognini-Bonelli (eds.), *Text and technology*. Amsterdam: John Benjamins, pp. 157–176.
- Louw, B. (2000). Contextual prosodic theory: Bringing semantic prosodies to life. In C. Heffer and H. Sauntson (eds.), *Words in context: A tribute to John Sinclair on his retirement*. Birmingham: University of Birmingham, pp. 48–94.
- Lyons, J. (1977). *Semantics* (Vol. 2). Cambridge: Cambridge University Press.
- Mahlberg, M. (2013). *Corpus stylistics and Dickens's Fiction*. London: Routledge.
- Michel, J-B., Shen, Y., Aiden, A., Veres, A., Gray, M., Google Books Team, Pickett, J., Holdberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. and Aiden, E. (2010). Quantitative analysis of culture using millions of digitized books. *Science* 311: 176–182.
- Moon, R. (2015). Multi-word items. In J. Taylor (ed.), *The Oxford handbook of the word*. Oxford: Oxford University Press, pp. 120–140.
- Moreno-Sánchez, I., Font-Clos, F. and Corral, Á. (2016). Large-scale analysis of Zipf's Law in English texts. *PLoS One* 11(1).
- Mouritsen, S. (2010). The dictionary is not a fortress: Definitional fallacies and a corpus-based approach to plain meaning. *BYU Law Review* (5): 1915–1980.
- Piao, S., Dallachy, F., Baron, A., Demmen, J., Wattam, S., Durkin, P., McCracken, J., Rayson, P. and Alexander, M. (2017). A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation. *Computer Speech & Language* 46: 113–135.
- Piao, S., Rayson, P., Archer, D., Wilson, A. and McEnery, T. (2003). Extracting multiword expressions with a semantic tagger. In *Proceedings of the ACL workshop on multiword expressions: Analysis, acquisition and treatment, at ACL 2003, the 41st annual meeting of the association for computational linguistics, Sapporo, Japan*. New York: Association for Computing Machinery, pp. 49–56.
- Plag, I. (2003). *Word-formation in English*. Cambridge: Cambridge University Press.
- Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.
- Rayson, P., Berridge, D. and Francis, B. (2004). *Extending the Cochran rule for the comparison of word frequencies between corpora*. 7th International Conference on Statistical Analysis of Textual Data. Available at: <http://eprints.lancs.ac.uk/12424/>.
- Saki. (2016). *The complete short stories of Saki*. London: Vintage.
- Sag, I.A., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In A. Gelbukh (ed.), *Computational linguistics and intelligent text processing: Proceedings of CICLing 2002*. Berlin and Heidelberg: Springer, pp. 1–15.
- Scott, M. (2009). In search of a bad reference corpus. In D. Archer (ed.), *What's in a word-list? Investigating word frequency and keyword extraction*. Oxford: Ashgate, pp. 79–92.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Solan, L. and Gales, T. (2016). Finding ordinary meaning in law: The judge, the dictionary or the corpus? *International Journal of Legal Discourse* 1(2): 253–276.
- Sorell, J. (2015). Word frequencies. In J.R. Taylor (ed.), *The Oxford handbook of the word*. Oxford: Oxford University Press, pp. 68–88.
- Steen, G.J., Dorst, A.G., Berenike Herrmann, J., Kaal, A.A. and Krennmayr, T. (2010). Metaphor in usage. *Cognitive Linguistics* 21(4): 765–796.
- Taylor, J. (ed.). (2015). *The Oxford handbook of the word*. Oxford: Oxford University Press.
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.
- Tweedie, F.J. and Harald Baayen, R. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32(5): 323–352.
- Weinreich, U. (1969). Problems in the analysis of idioms. In J. Puhvel (ed.), *Substance and structure of language: Lectures delivered before the linguistic institute of the linguistic society of America, Univ. of California, Los Angeles*. Berkeley: University of California Press, pp. 23–81.
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Ethnography

Piia Varis

Introduction

Digital humanities has, generally speaking, been mostly associated with practices such as creating large corpora, data mining large datasets and curating digital collections of cultural materials. Ethnographic research – despite its centrality as a ‘humanities’ approach, not to mention its long history as an important approach to the study of language – has not been widely seen as belonging to (the mainstream of) digital humanities research. At the same time, digital ethnography, the focus of this chapter, is increasingly used as an approach to understanding cultural formations, social relationships and linguistic and communicative practices in the digital age (see e.g. Murthy 2008; Underberg and Zorn 2013; Pink et al. 2016; Varis 2016a; Hou 2018; Maly 2018). Its role as somewhat of an outsider in digital humanities is at least to an extent due to the role of new technologies and media in digital humanities as defined from the *researcher’s* point of view, that is, as tools for making sense of people’s (linguistic) social lives; as defined by Terras (2016: 1644), ‘it is the role of the digital humanist to understand and investigate how computers can be used to question what it means to be human, and the human record, in both our past and present society’. Ethnography, on the other hand, has been less engaged in utilising new technological advances as tools for the researcher and more interested in finding out what new technologies and media mean for ordinary people and their (linguistic) practices.

Positioning digital ethnography as somewhat of an outsider is not to imply that digital humanities is somehow a clearly delineated approach to research – on the contrary, what exactly digital humanities stands for has been very much contested and debated over the years. The discussion on the relationship between computing and the humanities has also been partly conducted on a more general level, entailing broader questions regarding the issue whether (digital) humanities can be conceptualised as part of ‘social science’, or indeed, the humanities as ‘a subdomain of the social sciences’ (Rosenbloom 2012: np; on ethnography as an established social science perspective see e.g. Scott Jones and Watt 2010).

As for the precise definition of digital humanities, Terras (2016: 1640) states that ‘everyone using any computational method in any aspects of the arts and humanities is welcome!’ and points to digital humanities as a ‘rebranding’ of a lot of earlier computational work in the humanities, and as

a handy, all inclusive modern title which rebrands all the various work which has gone before it, such as Humanities Computing, Computing and the Humanities, Cultural Heritage Informatics, Humanities Advanced Technology.

Vanhoutte (2013: 147) points to a central question in these debates regarding the definition of the field, that is, that ‘the problem of self-presentation and self-representation remains with Digital Humanities’. Juxtaposing the earlier term for the field, ‘humanities computing’, with the more recent ‘digital humanities’, he writes:

Although Humanities Computing was a more hermetic term than Digital Humanities, it had a clearer purview. Humanities Computing relates to the crossroads where informatics and information science met with the humanities and it had a history built on the early domains of Lexical Text Analysis and Machine Translation. . . . Digital Humanities as a term does not refer to such a specialized activity, but provides a big tent for all digital scholarship in the humanities. . . . It still remains the question, however, whether well established disciplines such as Computational Linguistics or Multimedia and Game Studies will want to live under the big tent of Digital Humanities.

(Vanhoutte 2013: 144)

Interestingly, while Vanhoutte gives a very inclusive view of digital humanities (‘a big tent for all digital scholarship in the humanities’), he proceeds to mention disciplines which, although clearly representing ‘digital scholarship’ in the sense of engaging with digital material as their object of research, are not in a straightforward manner part of his definition of the field. Similarly, whether digital ethnography can be seen as a part of digital humanities is another matter of dispute.

The position assumed in this chapter is similar with Alvarado’s (2011: np) – that there is no need as such for a watertight definition of what ‘digital humanities’ stands for:

Instead of a definition, we have a genealogy, a network of family resemblances among provisional schools of thought, methodological interests, and preferred tools, a history of people who have chosen to call themselves digital humanists and who in the process of trying to define the term are creating that definition. How else to characterize the meaning of an expression that has nearly as many definitions as affiliates? It is a social category, not an ontological one.

Seen as a *social* category, or a ‘social undertaking’ ‘harbor[ing] networks of people (. . .) working together, sharing research, arguing, competing, and collaborating’ (Kirschenbaum 2010: 56), digital humanities can be defined as an affiliation or a point of identification. While ethnographers, specifically of the digital sort, can surely find space for identification there, the aim of this chapter, however, is not to try to persuade those who would not see ethnography as part of the digital humanities family to think otherwise. Rather, the aim is to discuss the possible contribution of ethnography as an approach in digitally mediated research, and as a specific approach to knowledge construction regarding linguistic

and communicative practices in the digital era. This means that the discussion here will not focus on, for instance, the use of ‘digital humanities’ tools (however one might want to define that) for ethnographic purposes – ethnography is methodologically flexible by default, so that is perhaps not necessarily such an interesting question to begin with. The aim here, then, is rather to highlight what is specific about ethnography and what kind of contribution it can make in the context of digital humanities research.

This chapter first briefly discusses the main tenets of the ‘pre-digital’ ethnographic tradition, building on the fundamental understanding that ethnography is not a method, but an approach (e.g. Blommaert and Dong 2010). The specific characteristics of digital ethnographic research will also be discussed and ethnography placed within a broader frame of recent developments in digital research. Next, critical issues in digital ethnography will be reviewed, with a particular focus on contextualisation. The specifics of digital ethnographic research and its particular strengths in describing and analysing communicative and linguistic digital practices will be further illuminated through a brief overview of current digital ethnographic research, and an analysis of ‘memic’ online communication (i.e. resemiotised signs involving memic-intertextual recognisability, in the case analysed here through a shared hashtag, see e.g. Shifman 2011; Varis and Blommaert 2015; Procházka 2018). Finally, future directions in digital ethnography will be addressed.

Background

Ethnography as an approach has a long history, with its roots in anthropological research. As ethnographers are interested in describing the reality ‘out there’, it is, of course, no surprise that it has gone digital, so to speak – everyday life and communication today are very much shaped by digital technologies, and it is thus natural that ethnography has taken ‘the digital’ as its object of study (see Coleman 2010; Varis 2016a for overviews). The ethnographic study of technologically mediated communication and practices is difficult to place under one label – different researchers, with different emphases, name their endeavours differently, which is why we have for instance ‘virtual ethnography’ (Hine 2000), ‘cyberethnography’ (Robinson and Schulz 2009), ‘discourse-centered online ethnography’ (Androutsopoulos 2008), ‘ethnography on the Internet’ (Beaulieu 2004), ‘ethnographic research on the Internet’ (Garcia et al. 2009) and ‘netnography’ (Kozinets 2009). The label favoured here is ‘digital ethnography’ (e.g. Murthy 2008; Varis 2016a). The motivation for that is that it is important to capture the fact that ethnography in our so-called digital age should not, and does not, focus only on what happens *online* but is also interested in the ways in which the online dimension is an integral part of the so-called *offline* sphere. Thus, ethnography prefixed as for instance ‘virtual’ or ‘online’ will only refer to a specific type of digital ethnography.

Christine Hine (2013: 8) suggests that ‘our knowledge of the Internet as a cultural context is intrinsically tied up with the application of ethnography’. While this may seem like a self-congratulatory statement, it is also true that ethnographers, with their interest in people’s lived reality, specifically aim at producing detailed, situated accounts of people’s linguistic and other social practices. In that sense, it seems realistic to assume that it is precisely ethnography that is well equipped for advancing our understanding of language use, communication and social relationships in digital contexts. With its focus on making sense in context, it also guards us against overly sweeping generalisations regarding language use in technologically mediated communication, such as ‘the language of emails’ (see Androutsopoulos 2008).

Established ethnographic approaches to the study of language in society have had an important role to play in advancing our understanding of language use and linguistic practices. As Page et al. (2014: 105) point out,

Within linguistics there has always been a strong strand of ethnographic work where linguists have studied the structure and uses of specific languages within a cultural framework, often grouped under the broader label of ‘Linguistic Anthropology’ (Duranti 1997; Ahearn 2010) and continuing up to the more recent ‘Linguistic Ethnography’.
(Creese 2010)

The more recent successor of these approaches, digital ethnography, which has appeared along with the increased digitalisation of life, is the focus of this chapter. Digital or not, the best way to describe what ethnographers do is, following Hymes (1996; also Velghe 2014), a ‘learning process’. Methodologically, ethnography is flexible – although mostly associated with fieldwork and interviews, it is impossible to definitively delineate what count as ethnographic methods: what makes most sense in explaining the issue at hand will be the method of choice for the ethnographer. The goal is to produce knowledge that is ‘ecologically valid’ (see Blommaert and van de Vijver 2013), and in that sense having pre-defined methods is not justifiable; ethnography is responsive to whatever arises in the field. This leads us to the critical issue of ‘context’ in digital ethnographic research, as well as more broadly in digital humanities.

Critical issues and topics

In the digital world, one of the most critical issues for ethnographers has to do with *context*, and context in two different senses: in terms of determining field sites and in terms of making sense of people’s social practices. The first issue has to do with the fact that ‘life online is not characterised by isolated, physically distinct spaces’ (Page et al. 2014: 105). Hyperlinked and connected as online environments are, an ethnographer often finds, for instance, that what happens on Instagram does not stay on Instagram: explaining digital practices often means having to account for materials appearing on many different platforms, or alternatively drawing what may seem like artificial boundaries around one’s field. As Page et al. (2014: 107) point out, ‘What constitutes a field site in networked, social media contexts is problematic, and in contrast to geographically-based field sites, may be “ambient” in nature which means that drawing those boundaries around the site can sometimes be tricky’.

Another issue regarding ‘context’ in digital ethnography is that in linguistic research, social media are perhaps not adequately accounted for as environments shaping interactions. This question will return later in the section ‘Sample analysis’, so it suffices to point out here that social media are not self-evident ‘contexts’, but just like any other social environment for interaction, are part of the ‘definition of the situation’ (Goffman 1959) in interactions. This also relates to what Page et al. (2014: 123) point to in their discussion of differences between ‘online’ and ‘offline’ research: ‘One of the things which can be different when researching the online world (and social media in particular) is that the researcher can already appear to be quite knowledgeable when starting the research’. Ethnographers can, of course, also be familiar with the offline environments that they study – in that sense it is difficult to draw a clear difference between ‘online’ and ‘offline’ research practices. However, it does seem to be the case that the features and affordances of online environments are

not extensively discussed in explaining for instance people's social media communications. This can be problematic in the sense that the end result might amount to decontextualised accounts of people's practices.

The risk of 'decontextualisation' is also present in the sense discussed by boyd and Crawford (2012) who provide a critique of what are essentially decontextualised 'patterns' in computerised (specifically big data) research:

Too often, Big Data enables the practice of apophenia: seeing patterns where none actually exist, simply because enormous quantities of data can offer connections that radiate in all directions. In one notable example, Leinweber (2007) demonstrated that data mining techniques could show a strong but spurious correlation between the changes in the S&P 500 stock index and butter production in Bangladesh.

(boyd and Crawford 2012: 668)

While digital humanities, of course, is about much more than correlations and big data, the issue of contextualisation in determining what exactly patterns are about is something to which ethnographers can make a contribution. Again, this issue will be revisited in the sections 'Sample analysis' and 'Future directions'.

Current contributions and research

Rymes' 2012 study of how YouTube videos are recontextualised and incorporated into local repertoires is a good example of how important it is to examine (re)uses and trajectories of semiotic materials online from one context to another and how they become to make sense, linguistically and otherwise, in different contexts. Due to the ease in mobilising and circulating them, semiotic material often has unpredictable trajectories, which need to be ethnographically examined to arrive at an understanding of how they fit into people's linguistic repertoires (see also Georgakopoulou 2013; Leppänen et al. 2014; Leppänen et al. 2017).

Another recent development in understanding people's language use in a digital world is multi-sited, multi-method ethnographic research which aims at understanding digital communication practices and language use within the wider sociolinguistic context. A number of such studies have already appeared, prime examples being the studies conducted in Copenhagen as part of the so-called Amager project (see Madsen et al. 2013 for an overview of the project; Madsen and Staehr 2014; Staehr 2014a; Staehr 2014b). This kind of multi-sited research helps simultaneously in introducing the 'online' as no longer a separate sphere of life, as well as clarifying what exactly is specific about the norms and affordances of online environments.

Another recent development, interesting for both ethnographers and (other) digital humanities researchers, is the combination of 'close reading' and 'distant reading' (Moretti 2005, 2013; see also Varis and van Nuenen 2017). A somewhat similar experiment is that of 'ethnomining' (Aipperspach et al. 2006), which combines data mining and ethnography. In combining 'distant' and 'close' reading, the idea is that 'computational results may be used to provoke a directed close reading' (van Nuenen 2016: 53). This means that patterns – and also outliers, for that matter, although the focus perhaps tends to be on patterns – established through 'data exploration' are then picked up for closer scrutiny (see van Nuenen 2016 for a discussion and examples). It can be assumed that such research will become more popular in the future; it seems likely that more and more interdisciplinary research will take place and that different types of (language) scholars will work together (more regarding this in the section 'Future directions'). This is, of course, also potentially interesting for ethnographers,

though it naturally involves a new way of approaching questions arising from ‘the field’; whether there is a difference between patterns and phenomena established assisted by computers instead of having been identified by the human ethnographer is an interesting epistemological and methodological question.

Auto-ethnographic and non-representational approaches are two further examples of recent contributions. As Bødker and Chamberlain (2016: 2) suggest, ‘autoethnographic renderings of mundane sociotechnical worlds are apt to elicit new perspectives of digitally mediated experiences’ and ‘new ways for representing and understanding forms of corporeal “felt-ness” of human-technology entanglements’ (ibid.: 12). Similar commitments can be detected in non-representational approaches concerned with ‘enlivening representation’ (Vannini 2015: 119), and

directed at making ethnographic representation less concerned with faithfully and detachedly reporting facts, experiences, actions, and situations and more interested in making them come to life. . . . In other words, by enlivening ethnographic representation I refer to an attempt at composing fieldwork as an artistic endeavor that is not overly preoccupied with mimesis.

(ibid.)

Such approaches going ‘beyond representation’ thus involve viewing ‘description not only as a mimetic act’, as ‘the lifeworld always escapes our quest for authentic reproduction’ (Vannini 2015: 122). Examples of work inspired by non-representational theory include Eslambolchilar et al. (2016) on the design of handheld technologies and Larsen (2008) on digital photography (see also Thrift 2008).

Finally, design ethnography (e.g. Crabtree et al. 2012) has been employed both in academic and commercial contexts, to understand users’ requirements and in developing and evaluating design ideas in the construction of computing systems. Crabtree et al. (2013: 1) advocate ‘in-the-wild’ approaches, which ‘create and evaluate new technologies and experiences *in situ*’, and ‘Instead of developing [design] solutions that fit in with existing practices, . . . experiment[s] with new technological possibilities that can change and even disrupt behavior’ (ibid.: 1–2). It should be noted that the work mentioned here represents specific approaches to, and definitions of, ethnography: Crabtree et al. (2012) delineate an ‘ethnomethodological approach to ethnography’ in design, and Crabtree et al. (2013) features a discussion of lab-based research as a ‘natural’ setting and seems not to include ethnographic research as ‘*in situ*’ research. In any case, ethnographic research on design and users’ experiences with computing systems seems to become ever more relevant considering their ubiquity in people’s everyday lives.

Main research methods

For ethnographers, digitalisation has indeed introduced opportunities for methodological reflection. As Christine Hine (2013: 9) points out,

The question is much more interesting, potentially, than whether old methods can be adapted to fit new technologies. New technologies might, rather, provide an opportunity for interrogating and understanding our methodological commitments. In the moments of innovation and anxiety which surround the research methods there are opportunities for reflexivity.

As mentioned earlier, as an approach, ethnography is methodologically flexible, and the epistemological commitment to producing knowledge that is ecologically valid can be seen as more important than the specific methods used for achieving that.

This also relates to the question of whether ‘on the screen’ research is adequate in explaining people’s linguistic and other behaviour or whether ‘off the screen’ research is always necessary. In this respect, while people’s offline realities do, of course, influence their online activities, there should perhaps not be too much of a focus on the online/offline issue. Online and offline research do have some differences, not least because of the differences in the ways in which data can be collected in these different environments. Online data are also ‘persistent’ (boyd 2008), which enables going back to sometimes very extensive ‘archives’ of communication. But the online/offline distinction may not be all that useful in thinking about people’s language use, at least if it is made into an issue of ‘authenticity’, that is, whether people behave online in ways that are somehow ‘inauthentic’, protected by the screen and sometimes also by anonymity. In pre-digital times, the ‘authenticity’ of data collected in classrooms for instance would not have been doubted for its ‘authenticity’ (i.e. questioning whether children there *really* talk the way they do in the classroom), but would have rather been interrogated for its specificity as it appears in that particular context. The same should apply to online spaces: they are social spaces in their own right, with their own practices and normativities. As long as there is an underlying assumption of online spaces somehow being separate from all the rest that language users do, the online/offline distinction remains unhelpful.

As online spaces are social spaces in their own right, both epistemologically and methodologically, therefore, it might be interesting to attempt to understand people’s media ideologies. Drawing on Silverstein’s (1979) notion of ‘language ideology’, Ilana Gershon (2010: 3) defined ‘media ideology’ as ‘a set of beliefs about communicative technologies with which users and designers explain perceived media structure and meaning. That is to say, what people think about the media they use will shape the way they use media’. Understandings of ‘proper’ or ‘correct’ uses of (social) media influence not only those designing the (social) media and their features (i.e. for their part shaping the ‘context’ in which interactions take place) but also those using them – what people see as ‘appropriate’ behaviour, linguistic or otherwise, within a medium will have effects on their communication. Consequently, as we will also see in the short sample analysis next, media ideologies are important in making sense of people’s digital media practices.

Sample analysis

Here, due to limitations of space, I will report on a small part of a broader endeavour to understand the phenomenon of the selfie (Varis 2016b). This section aims to illuminate the issue of *context* in relation to research on digital environments, in this case particularly social media. The issue discussed here first drew my attention in the literature on language use and self-presentation, specifically related to the ‘on my best day’ phenomenon (e.g. Baron 2010), but also from discussions with my students in classes where we analyse social media use and behaviour and what they wrote in their media diaries as part of their course work about their own social media use, for instance, about using Snapchat instead of other social media for certain purposes because ‘there are no expectations to feel attractive’ (Student media diary 2014).

‘Selfie’, the Oxford Dictionary Word of the Year in 2013, refers to ‘a photograph that one has taken of oneself, typically one taken with a smartphone or webcam and shared via social media’ (Oxford Dictionaries 2016). The selfie is one common form of presenting the self

on social media, and selfies ‘have become a genre unto themselves, with their own visual conventions and clichés’ (Marwick 2015: 141). Here, I will focus on one specific selfie phenomenon, namely the ‘ugly selfie’ – that is, selfies with what could be labelled as ‘transgressive aesthetics’ (Varis 2016b), that is, selfies where by definition the idea is for one to appear in one way or another as ‘unattractive’. Both so-called ordinary social media users, as well celebrities, have engaged in posting ‘ugly’ selfies of themselves and photos posted with the hashtag ‘ugly selfie’ can be found on many social media such as Instagram, Twitter and Tumblr. The related hashtag #uglyselfiechallenge became a popular trend in 2014 and entailed users inviting – or challenging – each other to post ugly selfies of themselves.

Judging by the number of photos posted on, for instance, Instagram, the ugly selfie is a very popular phenomenon; a search with the hashtag #uglyselfie yielded 56,315 posts (last checked 16 May 2016). The hashtag search is one of the most obvious ways of starting data collection for a viral or memetic social media phenomenon, and that is how the #uglyselfie research reported here also started. An immediate challenge that appears with hashtag data collection is that the numbers and the data that appear in a search may be misleading, as was also the case here: browsing through the images, an immediate finding was that not all of the photos tagged with #uglyselfie are by any means even selfies – they include pictures of cats, for instance, which we can probably safely assume are not selfies – and pictures of all kinds of everyday objects. There was also nothing particularly ‘ugly’ about all the pictures; rather, photos very much conforming to present-day standard understandings of beauty were to be found. Also, different memes and so-called motivational quotes appear with this hashtag, and interestingly, many of the quotes appeared to address the issue of ‘authenticity’ (e.g. ‘If I send you ugly selfies, our friendship is real’), which is actually what led to the examination of the theme discussed here. In other words, the manual shifting through and reading of the images and the attached hashtags started pointing towards a phenomenon which was not initially on the radar at all.

Here, for the purpose of illuminating the issue of ‘context’ in relation to social media research, I will be looking into the *discourse* surrounding the #uglyselfie phenomenon as it appeared on news media – a part of the project which aimed at further understanding the broader cultural frames within which the ugly selfie should be understood (see also Georgakopoulou 2016). This is interesting regarding norms of self-presentation on social media and understandings of ‘good’ and ‘appropriate’ use of social media regulating individual users’ self-presentation – in other words, ‘media ideologies’ as described by Gershon (2010). From the news discourse, we can discern some patterns in the way in which the ugly selfie is discursively constructed. For this sample analysis, I have looked more closely into seven online articles focusing specifically on the ugly selfie, the sources ranging from the *Teen Vogue* and *Seventeen* – aimed at younger audiences – to a site specifically aimed at women (*The Cut*), and *The New York Times*, *The Sydney Morning Herald*, *The Mail Online* and *The Telegraph*, that is, ‘mainstream’ newspapers.

In March 2014, the *Teen Vogue* featured a piece titled ‘The ugly selfie trend: How funny pics are turning social media on its head’. The title already points to transgression – the ugly selfie is ‘turning social media on its head’. The piece quotes what appear to be ‘ordinary’ social media users, such as 17-year-old Nina who is quoted as saying: ‘It might sound corny, but I think ugly selfies are pretty radical’, and

selfies are empowering because they allow us to decide for ourselves when we look our best. But so much of that is dependent upon how pretty I feel at a particular moment, which is only liberating to a certain degree! Ugly selfies reject that whole idea. Putting one up shows that how you look isn’t the most important thing.

Lilly, presented as a high school student from New Jersey, is another young woman quoted in the *Teen Vogue* article who, echoing Nina's view, reports that 'an intentionally ugly picture 'takes the pressure off of having to look your best'. In a similar vein, 13-year-old Ruby, quoted in a *New York Times* article from 21 February 2014, says about ugly selfies: 'It's fun', 'And it's just, like, so much work to make a good-looking selfie'.

Thus, while the selfie in general appears as empowering – giving one the power to present oneself through social media as one wants – at the same time, it is a genre of self-representation that places aesthetic pressures on these young women to look good. Editing and curating one's selfies also appears as a time-consuming activity. For instance, in the *Teen Vogue* article mentioned earlier, a social media user, Leah, is described as admitting to a 'previous obsession with cultivating a perfect web appearance' and 'spending hours combing through Facebook to untag any picture she deemed unflattering'.

Thus, while the selfie appeared as somehow empowering, as a practice of self-representation it is in fact largely presented in a very critical light. *The Mail Online*, for instance, in its ugly selfie article from 2013, refers to selfies as 'perfectly posed (and often highly edited)'. In a similar vein, the women's magazine *The Cut*, in a 2013 article described our time as 'a time when impossibly glossy, carefully constructed images – not just of models and celebrities, but of lifestyle bloggers, Instagram stars, and old acquaintances you now only see online – invade our lives daily'. *The Cut* also describes social media feeds as 'plastered with these very narcissistic "hot" pictures . . . almost as if [people were] scrutinizing their faces for perfection'.

Interestingly, the specific affordances of digital and social media feature prominently as encouraging the 'polished' nature of the selfie practice. Again, in *The Cut*, a poet and founder of a 'positive body image community' called *The Body Is Not An Apology*, Sonya Renee Taylor, is quoted as saying that 'digital cameras have created a world in which everyone is beautiful'. *Seventeen* magazine (2015), pointing to the same issue, states that 'In a world of photo-editing apps and filters, it can feel like there's tons of pressure to take the perfect selfie'. In a somewhat more alarming tone but again in a similar line of argumentation, *The Telegraph* mentions on 28 February 2014 that

Two weeks ago, Thailand's Department of Mental Health issued a warning about young-adult addiction to likes – and the deleterious effect un-liked selfies can have on like-addicts' minds. 'If they don't get enough 'likes' for their selfie as expected, they decide to post another, but still do not receive a good response. This could affect their thoughts. They can lose self-confidence and have a negative attitude toward themselves, such as feeling dissatisfied with themselves or their body.'

What is created here is a context in which social media users appear very much concerned with their looks to the extent of this being unhealthy, at least partly due to digital media with their affordances (such as 'like' buttons) which appear to feed the addiction. It is indeed a language of obsession and addiction that was used to describe all of this – obsession with looking good and with gaining 'likes' for instance as a form of acknowledgement.

It is in this context that the ugly selfie appears as transgression and a more positive genre of self-representation. *The Mail Online* calls the ugly selfie 'a backlash against the flawless selfie', *The New York Times* describes it as a 'a visual coup d'état' and *The Cut* says that 'ugly selfies . . . give permission to stop monitoring yourself for a moment'. And, finally, according to the *Teen Vogue*, 'sharing intentionally unattractive pics really is having a positive effect on many girls'. What exactly, then, is so radical and transgressive about the ugly selfie?

The Sydney Morning Herald (2016) explicitly makes a link between ugly selfies and feminist politics: 'When girls meet feminism in a toxic culture of appearance, ugly selfies are the result'. The ugly selfie is presented as a specifically *female* genre of self-representation, and *The Sydney Morning Herald*, for instance, emphasises its nature as a medium for intimate female relationships, that is, friendship between girls who send ugly selfies to one another. As one 14-year-old put it in *The Sydney Morning Herald* (2016) story: 'The closer the friend, the uglier the photo!' Apart from being an explicitly feminist act, then, the ugly selfie is presented as a medium for intimate relationships where one is able to expose themselves, show their 'real', unfiltered self through their so-called ugly photos.

Seventeen magazine (2015) indeed states that the initiators of the ugly selfie challenge 'want you to embrace your true self, and stop worrying about filters and likes'. A particular discourse of *authenticity* thus appears; *The Cut* article also specifically mentions that 'Speaking to women who have embraced the ugly selfie, these authenticity themes come up repeatedly'; the ugly selfie is a 'reminder to be genuine' as 'Conventional selfies are too posed'. Similarly, according to *The New York Times*, 'Amid the bared midsections and flawless smiles flashed all so often on the screen comes the explosion of the ugly selfie, a sliver of authenticity in an otherwise filtered medium' and 'a kind of playful slice of authenticity in an age where everything seems airbrushed to perfection'. *The Cut* also quotes an Australian feminist they interviewed as saying that the ugly selfie 'puts personality back into a superficial practice of self-documentation that has become devoid of anything real'.

This media discourse around selfies and ugly selfies alike seems to revolve a lot around whether these forms of self-representation are 'real' and 'authentic' or 'fake' and 'too polished'. The selfie and the ugly selfie are presented as requiring different amounts of work, which is also indicative in this respect: 'ordinary' selfies appear inauthentic, and are presented as requiring a lot of styling work, while the ugly selfie on the other hand appears as more effortless and spontaneous, 'more authentic', and 'real'. In making sense of this, the pre-digital work of Erving Goffman on the presentation of self is helpful. In *The Presentation of Self in Everyday Life*, Goffman states that

For many sociological issues it may not even be necessary to decide which is the more real, the fostered impression or the one the performer attempts to prevent the audience from receiving. The crucial sociological consideration, for this report at least, is merely that impressions fostered in everyday performances are subject to disruption.

(Goffman 1959: 72)

He thus further says that

We will want to know what kind of impression of reality can shatter the fostered impression of reality, and what reality really is can be left to other students. We will want to ask, 'What are the ways in which a given impression can be discredited?' and this is not quite the same as asking, 'What are the ways in which the given impression is false?'

(Goffman 1959: 72–73)

The appearance of the ugly selfie on social media, in Goffman's terms, appears to discredit and disrupt the normative everyday performance of the self through selfies. However, if we follow Goffman, then each form of self-representation is a performance in front of, and for, an audience. Also, interestingly for digital ethnographers and other students of digital communicative practices, the affordances of our digital media such as the filters, cropping

features and ‘like’ buttons available are ‘props’ in the Goffmanian vocabulary of stage performance, shaping the performance.

These built-in ‘defaults’, as van Dijck (2013: 32) points out, ‘are not just technical but also ideological manoeuvrings. . . . Algorithms, protocols, and defaults profoundly shape the cultural experiences of people active on social media platforms’. These coded structures, she (ibid.: 20) maintains, ‘are profoundly altering the nature of our connections, creations, and interaction. Buttons that impose “sharing” and “following” as social values have effects in cultural practices’.

At the same time, social media is intertwined with particular ideological understandings of authenticity. Gaden and Dumitrica (2015: np), in their discussion on authenticity in the context of social media, say that authenticity has become related to, first of all, ‘realism’ through the presentation of ‘detailed, mundane and personal content’, by making one’s ‘real self’ available, and, second, immediacy and the ‘live’ quality of the communication of the self that produces the feeling of authenticity’. They propose that

This view of the ‘real self’ enabled by immediate and regular updates [such as selfies] has to be understood against an emerging reconceptualization of traditional media as producing ‘delayed’ and ‘edited’ texts. Unlike such communication forms, the ‘authenticity’ of blogging [and other social media] appears as a result of the possibility to present your own self to the world on an ongoing basis, or ‘as it unfolds’.

Looked at this way, the media discourse around the ugly selfie phenomenon has to do with particular understandings of authenticity intersecting with media ideological, normative notions regarding the nature of social media communication. It appears to, first of all, prioritise offline selves and offline communication over online ones: the digitally unmediated self, presented as unfiltered, untouched, is the benchmark of authenticity against which the selfie is deemed as ‘fake’ and ‘polished’. This view, of course, happily ignores the fact that offline selves are also performances each with backstage for preparation and the front stage for the performance, in Goffman’s terms, and both the selfie and the ugly selfie are staged performances in the sense that the flow of everyday life is interrupted by making it fit a screen, to be posted with a specific online audience in mind.

Also, returning to the understandings of authenticity in relation to social media outlined earlier: the ugly selfie appears immediate, spontaneous, in the ‘now’, and as such adheres to media ideological understandings regarding the authentic self made available on social media. Unfiltered, it appears to show the prioritised offline self ‘as it supposedly is in the now’. The selfie, on the other hand, is delayed, edited and thus ‘inauthentic’, while contextually making sense, as it makes use of the features of social media such as cropping and filtering, contextual props readily available for the social and digital media user in their devices and interfaces and ready to be posted to sites with affordances encouraging quantification with ‘likes’ and ‘loves’ built into the interface, thus encouraging ‘likeable’, non-transgressive forms of self-representation.

In this understanding, we need to look into not only the discursive performances produced – such as selfies – but also the contextual affordances shaping them – such as filters and like buttons. It seems that thus far language and discourse scholars have not been very interested in these platform features shaping online interactions and self-representations – the focus has been on the discourse appearing in the social media environments, with social media as somehow a self-evident ‘context’ – with interfaces and their features having been more the interest of platform studies and cultural studies scholars.

This perspective is also useful in the sense that it allows us to start examining the norms of ‘selfieing’ without falling into the sometimes pathologising and hyper-critical discourse around the genre (cf. Georgakopoulou 2016) and see it as a genre in itself, being shaped as it is by the characteristics of digital social media and prevailing – often highly gendered – norms of self-presentation and aesthetics of the self, and, to borrow another term from Goffman, the often very complex participation frameworks in which they become part of our everyday interactions, everyday self-representation and everyday politics.

This leads to one final part of the broader frame, that is, the ‘intersection of the personal and the political’: as Tim Highfield (2016: 153) reminds us in his book *Social Media and Everyday Politics*,

The overlap between the personal and the political is extensive, on social media and in general. This is seen in both how the political is framed, around individual interests and experiences, and how the personal becomes politicized. Politics emerges out of the presentation of the mundane.

(2016: 39)

He continues (ibid.) to say that ‘By debanalizing platforms (Rogers 2013), we can treat the everyday and the mundane not as trivial but as the context for understanding the logics and practices of social media’. This is particularly what the everyday, mundane phenomenon of the ugly selfie helps us understand, by making visible the media ideological understandings underpinning the posting, circulation and reception of selfies.

From a digital humanities perspective, this type of analysis is interesting in terms of understanding people’s linguistic and communicative practices in digital environments. As I have argued elsewhere (Varis 2016a: 61, emphasis original), ‘attending to *both* language ideologies and media ideologies would perhaps provide powerful explanations for what people do and why they do it’. What is of interest for ethnographers here is the creation and attribution of meaning to particular semiotic artefacts, in this case, a hashtag and the semiotic material tagged (cf. Zappavigna 2011).

It is clear also from the #uglyselfie discussion that ‘the online environments studied cannot be taken as self-explanatory contexts, but need to be investigated for locally specific meanings and appropriations’ (Varis 2016a: 58). In practice, this may not be all that easy, but it seems that understanding the very context in which people’s semiotic practices appear is necessary for making sense of their practices. As van Dijck (2013: 20) has argued, the structures of social media ‘are profoundly altering the nature of our connections, creations, and interaction. Buttons that impose “sharing” and “following” as social values have effects in cultural practices’.

Future directions

Whether digital humanities as a field is defined through the tools employed or as a ‘social category’ as suggested by Alvarado (2011) earlier – or something else – remains a matter of debate. The question appears to be whether ‘the digital humanities tent is big enough’, and it remains to be seen, for instance, to what extent digital ethnographers and those ‘traditionally’ understood as fitting into the tent will find space for collaboration, and whether digital ethnographers will start incorporating computational methods more extensively in their research.

Lev Manovich (2011) has argued that

we want humanists to be able to use data analysis and visualization software in their daily work, so they can combine quantitative and qualitative approaches in all their work. How to make this happen is one of the key questions for ‘digital humanities’.

(p. 14)

There is merit to Manovich’s point, but perhaps there is another route to combining qualitative and quantitative (or, to bring in a different kind of co-operation: combining ethnography with ‘traditional’ digital humanities research) insights: namely teamwork, also of the interdisciplinary kind. As Manovich himself (2011: 14) points out, such a goal ‘requires a big change in how students particularly in humanities are being educated’. What Manovich is proposing thus seems like a tall order: training language and other humanities students in a different way, if this were to materialise, would take years to bear fruit. In the meantime, interdisciplinary collaboration makes the use of both qualitative and quantitative insights possible. Many, including Kirschenbaum (2010: 55), have identified digital humanities as ‘probably more rooted in English than any other departmental home’, and indeed English departments for instance seem like ideal humanities places for such collaboration to take place.

In a broader sense, this, of course, also has to do with epistemological questions and how humanities scholars produce knowledge. The so-called computational turn has not been without its critics and sceptics; Christian Fuchs (2017), for instance, has argued for a paradigm shift in the study of the Internet and digital/social media. Fuchs sees risks with the increasing use of, and funding for, computational research:

The danger is that computer science colonizes the social sciences and tries to turn them into a subdomain of computer science. There should certainly be co-operations between computer scientists and social scientists in solving societal problems, but co-operation is different from a computational paradigm. Turning social scientists into programmers as part of social science methods education would almost inevitably leave no time for engagement with theory and social philosophy, except if one radically increased the duration of study programmes. . . . An alternative paradigm that does not reject, but critically scrutinize [sic] the digital, is needed.

(Fuchs 2017: 47)

Fuchs’s concern has to do with theory: the idea he puts forward is that computational methods research is ‘undertheorised’, finds problems with solutions such as the one proposed by Manovich regarding the training of humanities students and scholars and therefore calls for an examination of the ‘digital’ in research. Whatever the way forward, it seems undeniable that ethnography has a role to play in digital humanities – or, in any case, within the humanities that interrogates the digital. As danah boyd and Kate Crawford (2012: 670) point out, ‘during this computational turn, it is increasingly important to recognize the value of “small data”. Research insights can be found at any level, including at very modest scales. In some cases, focusing just on a single individual can be extraordinarily valuable’. The contribution of digital ethnographers is also important for theory building and for making sense of the contexts in which people’s digital interactions take place. As we have seen from the #uglyselfie discussion, the digital (social media) contexts in which people’s communicative practices take place do play an important role in shaping the practices. Therefore,

as Blommaert and Rampton (2011: 10) point out, ‘the contexts for communication should be investigated rather than assumed’. Wherever digital humanities scholars go next, it is essential to consider that the ‘digital’ means not only new tools for analysis but also new types of contexts. These contexts cannot remain unscrutinised, as they as well as the meanings attached to them – or, in other words, media ideologies – shape the interactions and eventually the kind of ‘data’ researchers end up with. For an interrogation of these, digital ethnographers are well prepared.

Further reading

- 1 Blommaert, J. and Dong, J. (2010). *Ethnographic fieldwork: A beginner's guide*. Bristol: Multilingual Matters.
Blommaert and Dong's book is a good starting point for students and more advanced scholars alike; it provides an accessible outline of ethnographic research and is of particular interest to scholars of language.
- 2 Boellstorff, T., Nardi, B., Pearce, C. and Taylor, T.L. (2012). *Ethnography and virtual worlds: A handbook of method*. Princeton, NJ: Princeton University Press.
While focusing on virtual worlds and not written by a linguist, Boellstorff et al.'s handbook is valuable reading, as it discusses many aspects of ethnographic research of more general appeal for digital scholars.
- 3 boyd, d. (2008). *Taken out of context: American teen sociality in networked publics*. PhD Thesis, University of California, Berkeley.
Though not specifically focusing on language, danah boyd's PhD dissertation is a good introduction to the ethnographic study of digitally mediated life.
- 4 Hine, C. (ed.) (2013). *Virtual methods: Issues in social research on the Internet*. London: Bloomsbury.
Virtual Methods is a useful collection of chapters addressing different methods and issues related to researcher-researched relationships, the definition of field site and the online-offline dimension.
- 5 Page, R., Barton, D., Unger, J.W. and Zappavigna, M. (2014). *Research language and social media: A student guide*. London: Routledge.
Page et al.'s book focuses on social media in particular, but is a valuable resource in particular for students also for understanding research in the digital world in more general terms; its focus on language makes the book particularly interesting for students and scholars of language.

References

- Ahearn, L. (2010). *Living language: An introduction to linguistic anthropology*. Oxford: Wiley-Blackwell.
- Aipperspach, R., Rattenbury, T.L., Woodruff, A., Anderson, K., Canny, J.F. and Aoki, P. (2006). *Ethnominating: Integrating numbers and words from the ground up*. Technical report. Electrical Engineering and Computer Sciences, University of California at Berkeley. Available at: <https://www2.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-125.pdf>.
- Alvarado, R. (2011). The digital humanities situation. *The Transducer*, 11 May. Available at: <http://transducer.ontoligent.com/?p=717>.
- Androutsopoulos, J. (2008). Potentials and limitations of discourse-centred online ethnography. *Language@Internet* 5: art. 9. Available at: www.languageatinternet.org/articles/2008/1610.
- Baron, N.S. (2010). *Always on: Language in an online and mobile world*. Oxford: Oxford University Press.
- Beaulieu, A. (2004). Mediating ethnography: Objectivity and the making of ethnographies of the Internet. *Social Epistemology* 18: 139–163.

- Blommaert, J. and Dong, J. (2010). *Ethnographic fieldwork: A beginner's guide*. Bristol: Multilingual Matters.
- Blommaert, J. and Rampton, B. (2011). Language and superdiversity. *Diversities* 13(2): 1–21.
- Blommaert, J. and van de Vijver, F. (2013). Good is not good enough: Combining surveys and ethnographies in the study of rapid social change. *Tilburg Papers in Culture Studies*, Paper 65. Available at: www.tilburguniversity.edu/upload/61e7be6f-3063-4ec5-bf76-48c9ee221910_TPCS_65_Blomaert-ssssVijver.pdf.
- Bødker, M. and Chamberlain, A. (2016). *Affect theory and autoethnography in ordinary information systems*. Twenty-Fourth European conference on information systems (ECIS), Istanbul, Turkey, 12–15 June.
- Boellstorff, T., Nardi, B., Pearce, C. and Taylor, T.L. (2012). *Ethnography and virtual worlds: A handbook of method*. Princeton, NJ: Princeton University Press.
- boyd, d. (2008). *Taken out of context: American teen sociality in networked publics*. PhD Thesis, University of California, Berkeley.
- boyd, d. and Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society* 15(5): 662–679.
- Coleman, G.E. (2010). Ethnographic approaches to digital media. *Annual Review of Anthropology* 39: 487–505.
- Crabtree, A., Chamberlain, A., Grinter, R., Jones, M., Rodden, T. and Rogers, Y. (2013). Introduction to the special issue of “The turn to the wild”. *ACM Transactions on Computer-Human Interaction* 20(3): 1–4.
- Crabtree, A., Rouncefield, M. and Tolmie, P. (2012). *Doing design ethnography*. London: Springer.
- Creese, A. (2010). Linguistic ethnography. In L. Litosseliti (ed.), *Research methods in linguistics*. London: Continuum, pp. 138–154.
- The Cut*. (2013). *Ugly is the new pretty: How unattractive selfies took over the Internet*, 29 March.
- Duranti, A. (1997). *Linguistic anthropology*. Cambridge: Cambridge University Press.
- Eslamolchilar, P., Bødker, M. and Chamberlain, A. (2016). Ways of walking: Understanding walking's implications for the design of handheld technology via a humanistic ethnographic approach. *Human Technology* 12(1): 5–30.
- Fuchs, C. (2017). From digital positivism and administrative big data analytics towards critical digital and social media research! *European Journal of Communication* 32(1): 37–49.
- Gaden, G. and Dumitrica, D. (2015). The “real deal”: Strategic authenticity, politics and social media. *First Monday*. Available at: <http://firstmonday.org/ojs/index.php/fm/article/view/4985/4197>.
- Garcia, A.C., Standlee, A.I., Bechhoff, J. and Cui, Y. (2009). Ethnographic approaches to the Internet and computer-mediated communication. *Journal of Contemporary Ethnography* 38(1): 52–84.
- Georgakopoulou, A. (2013). *Small stories research and social media: The role of narrative stance-taking in the circulation of a Greek news story*. Working Papers in Urban Language & Literacies, Paper 100. Available at: www.academia.edu/6341450/WP100_Georgakopoulou_2013_Small_stories_research_and_social_media_The_role_of_narrative_stance-taking_in_the_circulation_of_a_Greek_news_story.
- Georgakopoulou, A. (2016). From narrating the self to posting self(ies): A small stories approach to selfies. *Open Linguistics* 2(1): 300–317. <https://doi.org/10.1515/opli-2016-0014>.
- Gershon, I. (2010). *Breakup 2.0: Disconnecting over new media*. Ithaca, NY: Cornell University Press.
- Goffman, E. (1959 [1990]). *The presentation of self in everyday life*. London: Penguin.
- Highfield, T. (2016). *Social media and everyday politics*. Cambridge, UK: Polity Press.
- Hine, C. (2000). *Virtual ethnography*. London: Sage Publications.
- Hine, C. (2013). Virtual methods and the sociology of cyber-social-scientific knowledge. In C. Hine (ed.), *Virtual methods: Issues in social research on the internet*. London: Bloomsbury, pp. 1–13.
- Hou, M. (2018). *Social media celebrity: An investigation into the latest metamorphosis of fame*. PhD dissertation, Tilburg University. Available at: https://pure.uvt.nl/portal/files/25867721/Hou_Social_23_05_2018.pdf.
- Hymes, D. (1996). *Ethnography, linguistics, narrative inequality: Towards an understanding of voice*. London: Routledge.

- Kirschenbaum, M.C. (2010). What is digital humanities and what's it doing in English departments? *ADE Bulletin* 150.
- Kozinets, R.V. (2009). *Netnography: Doing ethnographic research online*. London: Sage Publications.
- Larsen, J. (2008). Practices and flows of digital photography: An ethnographic framework. *Mobilities* 3(1): 141–160.
- Leinweber, D. (2007). Stupid data miner tricks: Overfitting the S&P 500. *The Journal of Investing* 16(1): 15–22.
- Leppänen, S., Kytölä, S., Jousmäki, H., Peuronen, S. and Westinen, E. (2014). Entextualization and resemiotization as resources for identification in social media. In P. Seargeant and C. Tagg (eds.), *The language of social media: Identity and community on the internet*. Houndmills: Palgrave Macmillan, pp. 112–136.
- Leppänen, S., Westinen, E. and Kytölä, S. (2017). *Social media discourse, (dis)identifications and diversities*. New York: Routledge.
- Madsen, L.M., Karrebæk, M. and Møller, J. (2013). *The Amager project: A study of language and social life of minority children and youth*. Tilburg Papers in Culture Studies, Paper 52. Available at: www.tilburguniversity.edu/upload/a0717d81-9116-45e8-85be-76466582b592_TPCS_52_Madsen-Karrebæk-Møller.pdf.
- Madsen, L.M. and Stæhr, A. (2014). *Standard language in urban rap: Social media, linguistic practice and ethnographic context*. Tilburg Papers in Culture Studies, Paper 94. Available at: www.tilburguniversity.edu/upload/1210cac3-2abf-42d7-84b5-f545916b4d15_TPCS_94_Stæhr-Madsen.pdf.
- Mail Online. (2013). *Forget the selfie, say hello to the uglie!* 24 October.
- Maly, I. (2018). New media, new resistance and mass media: A digital ethnographic analysis of the Hart Boven Hard movement in Belgium. In T. Papaioannou and S. Gupta (eds.), *Media representations of anti-austerity protests in the EU: Grievances, identities and agency*. New York: Routledge, pp. 165–187.
- Manovich, L. (2011). *Trending: The promises and the challenges of big social data*. Available at: <http://manovich.net/content/04-projects/067-trending-the-promises-and-the-challenges-of-big-social-data/64-article-2011.pdf>.
- Marwick, A.E. (2015). Instafame: Luxury selfies in the attention economy. *Public Culture* 27(1): 137–160.
- Moretti, F. (2005). *Graphs, maps, trees: Abstract models for a literary history*. London: Verso.
- Moretti, F. (2013). *Distant reading*. London: Verso.
- Murthy, D. (2008). Digital ethnography: An examination of the use of new technologies for social research. *Sociology* 42(5): 837–855.
- The New York Times*. (2014). *With some selfies, the uglier the better*, 21 February.
- Oxford Dictionaries. (2016). *Selfie*. Available at: www.oxforddictionaries.com/definition/english/selfie.
- Page, R., Barton, D., Unger, J.W. and Zappavigna, M. (2014). *Researching language and social media: A student guide*. London: Routledge.
- Pink, S., Horst, H., Postill, J., Hjorth, L., Lewis, T. and Tacchi, J. (2016). *Digital ethnography: Principles and practice*. London: Sage Publications.
- Procházka, O. (2018). A chronotopic approach to identity performance in a Facebook meme page. *Discourse, Context & Media*. <https://doi.org/10.1016/j.dcm.2018.03.010>.
- Robinson, L. and Schulz, J. (2009). New avenues for sociological inquiry: Evolving forms of ethnographic practice. *Sociology* 43(4): 685–698.
- Rogers, R. (2013). Debanalizing twitter: The transformation of an object of study. In K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann (eds.), *Twitter and society*. New York: Peter Lang, pp. ix–xxvi.
- Rosenbloom, P.S. (2012). Towards a conceptual framework for the digital humanities. *DHQ: Digital Humanities Quarterly* 6(2). Available at: www.digitalhumanities.org/dhq/vol/6/2/000127/000127.html.
- Rymes, B. (2012). Recontextualizing YouTube: From macro-micro to mass-mediated communicative repertoires. *Anthropology & Education Quarterly* 43(2): 214–227.
- Scott Jones, J. and Watt, S. (eds.) (2010). *Ethnography in social science practice*. Abingdon: Routledge.

- Seventeen. (2015). *Ansel Elgort and Jerome Jarre just started the best new selfie trend!* 30 January.
- Shifman, L. (2011). An anatomy of a YouTube meme. *New Media & Society* 14(2): 187–203.
- Silverstein, M. (1979). Language structure and linguistic ideology. In P. Clyne, W.F. Hanks and C.L. Hofbauer (eds.), *The elements: A parasection on linguistic units and levels*. Chicago: Chicago Linguistic Society, pp. 193–247.
- Stæhr, A. (2014a). The appropriation of transcultural flows among Copenhagen youth – The case of Illuminati. *Discourse, Context, and Media* (4–5): 101–115.
- Stæhr, A. (2014b). *Metapragmatic activities on Facebook: Enregisterment across written and spoken language practices*. Working Papers in Urban Language & Literacies, Paper 124, www.academia.edu/6172932/WP124_St%C3%A6hr_2014_Metapragmatic_activities_on_Facebook_Enregisterment_across_written_and_spoken_language_practices.
- The Sydney Morning Herald*. (2016). *Ugly selfies are all about girls trusting other girls*, 19 January.
- Teen Vogue*. (2014). *The ugly selfie trend: How funny pics are turning social media on its head*. 24 March.
- The Telegraph*. (2014). *Still taking selfies? It's "uglies" you should be perfecting, duh*, 28 February.
- Terras, M. (2016). A decade in digital humanities. *Humanities & Social Sciences* 7: 1637–1650.
- Thrift, N. (2008). *Non-representational theory. Space | politics | affect*. Abingdon: Routledge.
- Underberg, N.M. and Zorn, E. (2013). *Digital ethnography. Anthropology, narrative, and new media*. Austin: University of Texas Press.
- van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford: Oxford University Press.
- Vanhoutte, E. (2013). The gates of hell: History and definition of Digital | Humanities | Computing. In M. Terras, J. Nyhan and E. Vanhoutte (eds.), *Defining digital humanities. A reader*. Farnham: Ashgate, pp. 119–156.
- Vannini, P. (2015). Enlivening ethnography through the irrealis mood: In search of a more-than-representational style. In P. Vannini (ed.), *Non-representational methodologies: Re-envisioning research*. New York: Routledge, pp. 112–129.
- van Nuenen, T. (2016). *Scripted journeys: A study on interfaced travel writing*. PhD Thesis, Tilburg University.
- Varis, P. (2016a). Digital ethnography. In A. Georgakopoulou and T. Spilioti (eds.), *The Routledge handbook of language and digital communication*. London: Routledge, pp. 55–68.
- Varis, P. (2016b). *#uglyselfie as transgression*. Paper presented at the IGALA 9 (International Gender and Language Association) Conference: Time and Transition, City University of Hong Kong, 19–21 May.
- Varis, P. and Blommaert, J. (2015). Conviviality and collectives on social media: Virality, memes, and new social structures. *Multilingual Margins* 2(1): 31–45.
- Varis, P. and van Nuenen, T. (2017). The internet, language, and virtual interactions. In O. García, N. Flores and M. Spotti (eds.), *The Oxford handbook of language and society*. New York: Oxford University Press, pp. 473–488.
- Velghe, F. (2014). *"This is almost like writing". Mobile phones, learning and literacy in a South African township*. PhD Dissertation, Tilburg University. Available at: https://pure.uvt.nl/portal/files/4594667/Velghe_This_is_almost_03_12_2014.pdf.
- Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on Twitter. *New Media & Society* 13(5): 788–806.

Mediated discourse analysis

Rodney H. Jones

Introduction

Change the instruments, and you will change the entire social theory that goes with them.

– Latour (2009: 9)

The Japanese conceptual artist On Kawara (1932–2014) made an art of measuring his life. Among his most famous works is a collection of paintings which he made daily over the course of 48 years showing only the date. He called these timestamp pictures ‘self-portraits’. Other projects involved sending postcards to people every day telling them what time he got up, recording the path he took each day in red ink on a map of whatever city he happened to be in and making meticulous lists of everyone he met as he went through his life. Kawara has become somewhat of a cult hero to a group of people who are also obsessed with recording and counting the minutiae of their daily lives: the loose configuration of self-trackers and ‘body hackers’ (Dembrosky 2011) that have come to be associated with the ‘Quantified Self Movement’ (Abend and Fuchs 2016; Wolf 2010).

One of these is Cristian Monterroza,¹ a student at New York University (NYU) who shared his story in the form of a ‘show and tell talk’ at the New York City Quantified Self Meetup in November 2012. A ‘show and tell talk’ is a genre particular to Quantified Self Meetups, a personal narrative which combines elements from confessional stories (such as those told at Alcoholics Anonymous meetings) with slick professional PowerPoint presentations (like those shown at TED Talks) (Jones 2013a, 2016). Such talks are organised around three questions: What did I do? How did I do it? and What did I learn? ‘About a year ago,’ Monterroza begins, ‘I didn’t like the way my life was going. I felt like a gradual slip taking over. And I couldn’t quite put my finger on it. So, this quote came to mind: Insanity is doing the same thing over and over again and expecting different results’. He goes on to explain how, inspired by the work of On Kawara, he embarked on an ambitious project to track where he was and what he was doing every minute of his life, eventually developing an iPhone app which helped him to record where he went, when and how long he stayed

there and then convert these data into charts and graphs. All of this analysis, however, left him vaguely dissatisfied. Finally, he realised that his real insights were not coming from aggregating and quantifying these moments, but from looking at them *one at a time*, from looking at particular timestamps and particular global positioning system (GPS) locations and identifying in them ‘notable moments’, moments when he met someone, or tried something for the first time, or learned something. ‘Just like On Karawa proved,’ he concludes, ‘a small timestamp can mean a lot more, I’ve started realising that quantification, it’s not all about the numbers, numbers are very important, but I think we can aspire to something higher. I think it’s also about the perspective that it allows you to gain’.

In this chapter, I will use the stories of self-quantifiers like Christian Monterroza to discuss how technology can affect our study of the humanities and the way the humanities can offer insights into our encounters with technology. The theoretical framework that will form the basis of this discussion is *mediated discourse analysis* (Norris and Jones 2005; Scollon 2001), an approach to discourse which focuses on how the semiotic and technological tools we use to interact with the world serve to enable and constrain what we can know and who we can be. Mediated discourse analysis sees the analysis of texts and technologies as occasions for understanding how human social life is constituted and how it might be constituted differently through the exercise of human agency that can come as a result of a heightened awareness of the *mediated* nature of our experience of reality. For self-quantifiers like Christian Monterroza, mediated discourse analysis provides a way to examine how the timestamps and iPhone apps and ‘show and tell talks’ and social identities he uses to organise his search for his ‘real self’ act to determine the kind of self he is able to find and what he is able to find out about it. For researchers in the field of digital humanities who are likely to be readers of this book, it provides a way to reflect on how the tools we use to transform language, history and art into data also end up transforming what we consider language, history and art to be and who we consider ourselves to be as researchers. It reframes key questions about what we regard as knowledge and the nature of research as questions about the nature of *mediation* and the ways in which tools affect our actions, our perspectives, our values and our identities, and it reframes the mission of scholars in the digital humanities as not just a matter of using software to analyse texts but of analysing *how people use software* and how it changes the way they interact with texts.

Mediation

What distinguishes mediated discourse analysis from other approaches to discourse is its concern not with discourse *per se*, but with the social actions that discourse, and the technologies we use to produce it, make possible. Drawing on insights from Soviet psychologist Lev Vygotsky (1962), mediated discourse analysts see all social actions as mediated through one or more ‘cultural tools’. These tools might be physical tools, such as hammers or iPhones or laptop computers, or they might be semiotic tools, such as languages, images, charts and graphs, genres or styles of speaking or writing (including such ‘new’ genres as ‘databases’ and ‘corpora’). What we mean when we say actions are mediated through such tools is that when they are appropriated by particular social actors in particular situations, they make certain kinds of actions more possible (and other kinds of actions less possible). The reason for this is that different tools come with different ‘affordances’, a term which the psychologist James J. Gibson (1986) used to describe the potential that things in the environment, including technologies, have for serving as tools to perform certain actions. Affordances are, as (Gee 2014: 16) puts it, ‘what things are good for, based on what a user can do with them’.

This notion of ‘what things are good for’, however, is quite complex. What makes Cristian Monterroza’s iPhone ‘good for’ tracking his location or the narrative structure of a ‘show and tell talk’ ‘good for’ explaining what he did and what he learned has to do not just with the technical or structural properties of these tools but also with their histories, the various conventions of use that have built up around them, the ways they have come to be associated with particular social practices and particular ‘communities of practice’ (Lave and Wenger 1991) and the way they have come to be part of Cristian Monterroza’s ‘historical body’ (Jones 2007; Nishida 1958; Scollon and Scollon 2004), the practices, beliefs, knowledge, competencies and bodily dispositions that he has accumulated throughout his life.

Tools don’t just allow us to do certain things; they also channel us into particular *ways of being*, particular *ways of thinking* and *particular ways of relating* to other people (Jones and Hafner 2012). Cristian Monterroza’s iPhone does not just allow him to track his location and movements but also leads him to perceive his location and movements as somehow connected to his well-being. Apps that share his location with other people reinforce this thinking by helping to make ‘checking in’ a social practice and making Cristian’s friends with whom he shares his location feel that they are part of a ‘community’ of location trackers. The conventional structure of the ‘show and tell talk’ that Cristian gives at the New York Quantified Self Meetup – built around the three questions: What did I do? How did I do it? and What did I learn? – not only gives him a way to mentally organise what happened to him; it also gives him a way to ‘fit in’ with the other people in this community – other young, ‘tech-savvy’, financially secure people for whom ‘learning things’ and, more importantly, ‘talking about what you have learned’, are highly valued practices.

In other words, within the technological affordances of the various kinds of ‘equipment’ Cristian uses to ‘know himself’ there are already embedded certain *ways of knowing*, a certain kind of *self to be known* and a certain kind of *social order* in which these ways of knowing and these selves can be traded as ‘symbolic capital’ (Bourdieu 1992). The affordances of cultural tools have consequences, in that they don’t just allow us to do certain things, but they also shape the things that we do and that we *value* doing and the ways that we enact our identities and our social relationships with others. As Father John Culkin (1967: 70), in an oft-quoted commentary on the work of Marshall McLuhan, puts it: ‘We become what we behold. . . . We shape our tools and thereafter our tools shape us’ (see also Latour 2009). This shaping occurs on at least three dimensions: mediation is an *ontological* process, in that it shapes what we regard as real, how we divide up our world into concepts and categories, and the ‘mode of existence’ that we confer upon objects and ideas and that we claim for ourselves; it is an *epistemological* process, in that it shapes how we can know these objects and ideas and how we evaluate the validity of that knowledge; and it is an *axiological* process, in that it shapes what we value, what we consider ‘good’ or ‘moral’ and how we think societies should be organised and how people should treat one another.

But to say that tools shape their users, that for all of the benefits mediation brings it also brings limitations, that affordances are also always to some degree *constraints*, only tells half the story. At least it only represents half of the story that Cristian Monterroza tells, for the real point of his story is not what he learned from quantifying his daily movements, but what he learned about the *limits* of quantification. It was not enough, he concludes, to count how many times or how long he spent in different places or what he was doing there or how he felt when he was doing it. He also had to consider the *quality* of each individual moment as unique, unrepeatable and ‘notable’. What affords this ‘higher’ perspective is Cristian’s ability to combine quantification with *qualification*, to temper the objectivity of ‘hard data’ with a more subjective, hermeneutic perspective.

This, I would like to argue, is the really important thing about mediation; that as much as our tools shape us, this shaping is not determinative. It's not determinative, first, because human beings have the ability to be reflective about their tool use and to *imagine* different kinds of outcomes under different circumstances, and second, because we almost always have more than one tool available to us, and are able to combine different tools in ways that allow us to play the constraints of one tool off against the affordances of another. To return to Gee's formulation of Gibson's theory of affordances which I cited earlier, affordances come not just from what tools allow us to do but also from what we are able to do with those tools. In what follows I will discuss this double-edged quality of mediation as it relates to self-quantifiers who are trying to find 'self-knowledge through numbers', and, at the end of this chapter, relate this to the situation of scholars in the digital humanities, who are also involved in seeking knowledge about 'humanity' through numbers. I will consider, from the perspective of mediated discourse analysis, the 'digital' part of digital humanities: the potential of digital technologies to shape the ontological, epistemological and axiological dimensions of our research. But I will also consider the 'humanities' part of digital humanities, the capacity that we as scholars – and as humans – have, as Cristian Monterroza puts it, to 'aspire to 'something higher' by combining technologies in creative ways and by making particular aspects of our data 'notable' through hermeneutic processes. Along with Hayles (2012) and other scholars in the field (see for example Burdick et al. 2016; Rosenbloom 2012; Simanowski 2016), I will argue that the digital humanities should not just be concerned with how technology can help us to understand the humanities, nor with how the humanities can help us to understand technology, but with the 'relational architecture' (Rosenbloom 2012) between different ways of looking at and experiencing the world and how we can begin to trace the contours of that architecture.

The data I will draw on in this discussion come from a three-year-long ethnographic study of the Quantified Self Movement, involving attending 'Meetups' and conferences in five different countries, collecting and analysing 73 'show and tell talks', interviewing 'quantified selfers' and using various technologies to quantify my own behaviour and physical responses (such as heart rate, weight, sleep patterns) (see Jones 2013a, 2013b, 2013c, 2015a, 2015b, forthcoming). The study made use of the principles and methods of *nexus analysis* (Scollon and Scollon 2004), the methodological component of mediated discourse analysis which allows researchers to explore the complex ways that discourses and technologies circulate through communities through a combination of ethnographic observation and the close analysis of texts and interactions. In the final section of this chapter I will briefly describe the methodology of nexus analysis and discuss how it can be combined with other approaches to the digital humanities.

Mediation as ontology

Not content with just using his mobile phone to track his location and movements as Cristian Monterroza did, Miles Klee (2014) put it to work tracking his sex life. One of the apps he used was called *Love Tracker*, which includes a function that allows users to measure the duration of their lovemaking. 'Here's the thing, though,' he writes, 'figuring out when to start the timer is a nightmare. The app boasts a stopwatch function, but was I meant to flick it on as soon as I lunged toward my wife's side of the couch and, by extension, reached second base? Or should I start it when, after 20 seconds of making out, she realized that I wasn't going to leave her alone until she shut me down or acquiesced to my clumsy advances?'

Another aspect of the app that unnerved Klee was the fact that it would not allow a user to claim to have had sex for longer than 23 minutes.

Love Tracker is just one of many apps that provide different ways of measuring lovemaking. Another one, *Blackbook*, allows you to rate your sexual partners on a scale of 1 to 10, and another, *Sex Stamina Tester* allows you to count how many sexual partners you have had and how many times you have had sex with them and allows you to compare your score with other users, and yet another, *Spreadsheets*, uses the sound and motion sensors in your phone to measure the intensity of your sexual encounters in terms of the volume of your groans and ‘thrusters per minute’. What is common to all of these apps is that they do not just record and measure a particular phenomenon, but they also operate to *define* that phenomena in terms of what they are able to record and measure, whether it be frequency, duration, movement or the intensity of sounds as captured by the microphone of an iPhone (Jones 2015b). In other words, these apps don’t just ‘count’ different aspects of sexual behaviour; they also shape ‘what counts’ as sex. Herein lies the crux of mediation as an ontological process: the fact that technologies serve to determine how we define the objects of our study, usually based on the aspects of those objects that are ‘legible’ to whatever mediational means we are using. In other words, any attempt to understand something ends up to some degree creating the thing that we wish to understand.

Linguistic anthropologists Charles Bauman and Charles Briggs (1990) capture this idea that we create a phenomenon by separating it out from surrounding phenomena with their concept of *entextualisation*. ‘Entextualisation’, they write, is ‘the process of rendering discourse extractable, of making a stretch of linguistic production into a unit – a text – that can be lifted out of its interactional setting’ (1990: 73). Bauman and Briggs were mostly concerned with the process through which discursive phenomena – specifically, oral performances – are rendered ‘decontextualisable’. As I have argued in previous work (Jones 2009), however, entextualisation can be seen not as just a way of lifting discourse (speech, written words, images) out of their original context but as the primary mechanism through which we *capture* our actions and experiences and turn them into texts. This may be accomplished, for example, by describing a sexual encounter in a diary, filming or photographing it, logging it in a database based on a set of pre-set descriptors or recording it using the sensors in an iPhone. In all of these instances we are performing what is essentially an ontological activity by choosing (via whatever mediational means happen to be available to us) which aspects of the phenomenon to capture and to assign the label ‘sex’.

All forms of research make use of ‘technologies of entextualisation’ (Jones 2009: 286), mediational means which inevitably channel us into what Halpern (2014) refers to as ‘object oriented thinking’, the tendency to treat complex phenomena as more or less concrete objects. This is quite obvious when we employ instruments that ‘operationalise’ abstract concepts, such as when we use psychometric questionnaires to measure things like motivation or creativity. It may be less obvious when we observe things that seem already quite concrete, as when we use a microscope to peer at the wing of a fly, a telescope to gaze upon a planet or a computer programme to count lexical or grammatical features within a text. But in all of these cases the instruments that we are using are not just measuring wings and planets and texts, they are *creating* them by extracting them from some larger entity or stream of phenomena of which they are part. Latour (1999) makes a similar point in his examination of dirt in the Amazon, arguing that the only way to study soil is to extract and *abstract* it from the earth of which it is a part. The digital artist John Cayley (2016: 77) (see also Kitchin and Dodge 2011) insists on referring to the information we gather through our technological instruments not as ‘data’ but as ‘capta’, because it represents only ‘the captured

and abducted records' of that portion of human life that we are able to record with our technologies rather than our 'full (phenomenological or empirical) experience of the world'.

Mediation – and the processes of entextualisation that it makes possible – does not just create the objects of our study; it also creates *us* as certain kinds of observers or authors of those objects. In a focus group interview I conducted with self-quantifiers at a QS conference in Amsterdam, one participant mused:

There is something about every app that tries to paint you as a certain kind of person. So for example with Fitbit, the app already tells you that you're an active person because it gives you numbers about your physical activity, and the Lift app tells you that you are someone who is social, who wants to share goals with friends, because it gives you graphs comparing you with them. So, you need to decide if you want to accept the identity that the app creates.

This realisation that the act of observing a particular phenomenon acts not just to construct the phenomenon but also to construct us as observers is one of the reasons Miles Klee (2014) decided to give up his quest to quantify his sex life, concerned that, as he puts it, 'Technology had transformed me from a considerate lover into a number-crunching monster'. The problem here is not with number-crunching *per se*, but with the fact that an inevitable consequence of creating an object of study is that we assert our independence from that object. Sometimes such separation is enormously useful, providing a more objective perspective on the phenomenon we wish to study. In the context of the Quantified Self Movement, in fact, one of the aims of turning oneself into an object of study is to be able to see 'a version' of the self that we had not seen before. In the context of one's sex life, however, such separation might turn out to be counterproductive, especially from the point of view of one's partner. In the case of Miles Klee, he found himself beginning to orient less towards the phenomenon of sex or the person with which he was having it and more towards the technologies of entextualisation he was using to define and measure his sex. 'I noticed a funny anxiety emerging', he writes. 'Somehow I wanted to impress the apps, these nonsentient pieces of haphazardly designed software . . . alarmingly, I felt that I had something to prove to the sex-tracking apps'.

Comments like this tell us as much about ontology as they do about quantification, reminding us that at its core ontology is essentially a matter of performativity (Butler 1993): we call things into existence through entextualising them, and after multiple iterative acts of entextualisation, we come to regard the texts that we have produced as real. Belliger and Krieger (2016: 1) argue that 'personal informatics and body tracking is a *performative enactment of the informational self* (emphasis mine), a way of bringing ourselves into existence as particular kinds of beings, selves which are the product of particular historical circumstances and the 'technologies of the self' that they have made available to us (Foucault 2003). This idea of the informational self, however, is not new. In his 1738 *Treatise of Human Nature*, David Hume (1985) took up the question of why it is so difficult for us to know ourselves. One answer he offered was our tendency to confuse our ideas about identity with actual identity, exemplified in the fact that when an object changes very quickly we are apt to assign it a new identity, while when it changes slowly we are apt to feel that its identity has not fundamentally changed. In other words, we rely on our sense of what Hume calls 'numerical identity' to come to conclusions about what he refers to as our 'specific' (or qualitative) identity (see also Abend and Fuchs 2016).

But the object-oriented performativity evident in the practices of the Quantified Self Movement (or, for that matter, in the practices of biologists, computer scientists or digital

humanities scholars), harkens back to an even more fundamental ontological debate begun by Spinoza and Descartes almost a century before Hume about the nature of the self, specifically whether or not is possible to separate the body from the mind. In this debate, the dualistic ontology promoted by Descartes ([1641] 1986), in which a clear line is drawn between cognitive and corporeal selves, has seemingly won out, dominating social science (and, increasingly, humanities) research. The more monastic view of Spinoza ([1677] 1996), however, which views corporeal and cognitive dimensions as inherently *entangled*, has also gained currency, especially in the work scholars such as Butler (1993), Irigaray (1985), and Deleuze and Guattari (1987). What mediated discourse analysis adds to this debate is the framing of this philosophical question is a question about *mediation* and the effect of technologies on the way we *entextualise* the world and draw boundaries between external objects and ourselves (Scollon 2001).

Mediation as epistemology

Not all of the members of the Quantified Self Movement use high-tech gadgets to mediate their search for self-knowledge. Some, like Amelia Greenhall,² make use of more analogue mediational means. Greenhall has made a star chart for herself of the type one sees on the walls of kindergarten classrooms; whenever she completes an activity which she thinks is beneficial to her well-being, she gives herself a gold star. By looking at the number of stars she has accumulated over certain period of time, she can get a sense of how she is doing. The most beneficial thing about creating data in this way, she reasons, is that it motivates her to improve by ‘making small progress visible’. This low-tech measurement of performance is really not so different from the more high-tech means employed by people like Sky Christopherson,³ who uses digital sensors to measure things like movement, sleep, diet and heart rate and turn them into charts and graphs to improve his athletic performance; these charts and graphs serve the same function as gold stars: they make Christopherson’s progress visible, and once it is made visible, it is also made *actionable*: ‘If I can see a number attached to something, so I can break a record or something’, he says, ‘then I can get behind it. . . . If it can be measured, it can be improved’.

Greenhall’s and Christopherson’s experiences illustrate the epistemological dimension of mediation, the fact that the way we are able to know about the world depends upon the mediational means (whether they be gold stars or sophisticated graphs) that we use to represent it. From the point of view of mediated discourse analysis, just as ontology is essentially a discursive process of ‘capturing’ reality and turning it into a text, epistemology is a process of making phenomena ‘tangible’ (and thus ‘knowable’) by transforming it into different semiotic modes to which we give the name ‘data’. Understanding epistemology as a matter of the discursive processes which our technologies allow to engage in changes the way we think about ‘data’. Data do not exist independent of these processes; there is no such thing as ‘raw data’ (Gitelman 2013). Data are, as Engel (2011: np) puts it, ‘an epistemological alibi’, and excuse for adopting a particular way of knowing the world.

Whereas the main discursive process implicated in the ontological dimension of mediation is entextualisation, the extraction of phenomena from the stream of existence, the discursive process implicated in its epistemological dimension is that of *semiotisation*, or, as Iedema (2001) refers to it, *resemiotisation*. Semiotisation is the process of ‘translating’ whatever has been extracted through our instruments into different forms of semiosis (words, numbers, shapes, colours) in order to render it ‘meaningful’. Just as different technologies bring with them different affordances in terms of what aspects of reality they are

able to capture and extract, different semiotic modes bring with them different affordances and constraints when it comes to the different kinds of meanings that can be made with them (Kress 2009). In other words, different semiotic modes fundamentally embody different *theories of knowledge* (Latour 2009). We might, for example, resemiotise a collection of ‘capta’ as a bar graph, or a pie chart, or a table of numbers, or a written narrative, or a painting – each of these different forms of semiosis making it possible to know the phenomenon in different ways.

Resemiotisation is not just a valuable tool to think with; it also affects the ways we are able to *invest* in the knowledge that we create and what we are able to use that knowledge to do, a fact evidenced in the earlier examples of Amelia Greenhall’s gold stars and Sky Cristopherson’s high-tech charts and graphs. In an earlier article on the different ways self-tracking operates by *semiotising* phenomena (Jones 2015a), I discussed how self-tracking apps like Nike+ and Hydrate (an app for keeping track of how much water you drink) make use of different semiotic modes to construct both knowledge and ‘ways of knowing’ for their users. These processes of (re)semiotisation are extremely useful when it comes to monitoring our health and attempting to alter our behaviour, because they afford new ways of *experiencing* that behaviour. Numerous studies (see for example DiClemente et al. 2000; Frost and Smith 2003) have shown that presenting information about their health to patients in different modes; or on different timescales; or in different contexts can help them to think (and act) differently about things like diet, smoking and exercise by making available to them new practices of ‘reading their bodies’ (Jones 2013a: 143). The benefits of reading one’s body through self-tracking apps are, in many ways, similar to the benefits of the new practices of ‘distant reading’ (Moretti 2013) that have been applied to corpora of literary works, historical documents and other linguistic behaviour in the digital humanities. Being able to aggregate large amounts of data and analyse them using computational tools and processes has enabled us to read texts in whole new ways and find patterns and connections that we would not otherwise have been able to discern.

At the same time, there are also drawbacks associated with these processes of semiotisation. First, they inevitably constrain the way users think about health and health behaviour within the semiotic parameters of whatever modes the mediational means they are using are able to produce. Phenomena like well-being, sexual pleasure and ‘health’ – which are basically ‘topological’ phenomena subject to all sorts of subtle gradations – are often reduced to ‘typological’ phenomena – expressed as discrete measurements or categories (Lemke 1999). While this reductionism serves the purpose of making phenomena more ‘legible’, it also inevitably distorts them. Another drawback is that semiotisation has a tendency to lead us to reify knowledge: knowledge comes to take on a more and more solid material existence. Iedema (2001) uses the example of a building project to discuss how processes of resemiotisation involve not just changing meanings from one semiotic mode to another but often involve the progressive materialisation of those meanings: ideas about a new hospital wing spoken in the context of a meeting are resemiotised in written form in minutes of the meeting, and later in graphic form in architects’ blueprints, and finally in the form of bricks and mortar after the hospital wing has actually been built. We do the same thing with the phenomena we study – whether they be linguistic phenomena or bodily phenomena; we build edifices out of them through processes of entextualisation and resemiotisation and then sometimes mistake these edifices for the phenomena themselves.

This capacity for reification seems to be particularly strong when it comes to the affordances of digital technologies that help us to turn data into visualisations (charts, graphs, infographics, etc.). Data visualisation has become a central part of many contemporary

scientific endeavours, as well as contemporary journalistic practices, and has more recently been enthusiastically taken up by scholars in the humanities (Manovich 2013; Wouters et al. 2012). In all of these fields, visualisations have been praised for their ability to make new relationships between data visible and to produce new spaces for speculation. At the same time, as Halpern (2014: 22–23) points out, visualisation ‘invokes a specific technical and temporal condition and encourages particular practices of measurement, design, and experimentation’. Visualisations bias their readers to pay attention to certain meanings over other meanings and gradually train them in particular ways of paying attention to the world and information about it (Jones 2010). The most dramatic effect of the rise of visualisation as a way of experiencing data, according to Halpern, has been a shift in dominant ways of knowing from more discursive processes of ‘reason’, exemplified in written arguments, to more quantitative processes of ‘rationality’ in which knowledge is a matter of ‘facts’ derived from measurement and comparison. Drucker (2016: 63–64) argues that the difference between representing data verbally and visually is essentially an epistemological difference, writing:

When you realize that language has many modalities – interrogative and conditional, for instance – but that images are almost always declarative, you begin to see the problems of representation inherent in information visualizations. They are statements, representations (i.e. Highly complex constructions and mediations) that offer themselves as presentation (self-evident statements). This is an error of epistemology, not an error of judgment or method.

Another dramatic effect of the new popularity of data visualisation has been to create new ways for people to emotionally ‘invest’ in knowledge by highlighting its ‘aesthetic’ dimension. Data have become more and more ‘beautiful’ (Halpern 2014), especially in the popular press, but also in the humanities where the ‘beauty’ of data visualisations has come to act as a surrogate for the beauty of the literary or artworks which they represent. Data visualisations allow scholars in the digital humanities to make their endeavours seem simultaneously more aesthetic and more ‘scientific’. The notion of ‘beauty’ has come to imply value, usefulness and even ‘credibility’.

Mediation as axiology

Jon Cousins⁴ was depressed. After weeks of battling the bureaucracy of the mental health system in Britain, he decided to take matters into his own hands by developing a website which he used to record his mood on a daily basis and share his measurements with his friends. ‘When I first started measuring’, he says, ‘my mood went up and down, up and down, but then when I started to share with my friends, look . . .’ and he draws his finger across a sustained line at the top of his graph. What he learned from this, he said was, ‘if you share your mood with friends, your mood goes up. . . . It’s the sharing bit that seems to be the powerful bit’.

J. Paul Neely⁵ was also interested in using quantification to optimise his happiness, but he went about it in a different way. He decided to try an experiment during a gathering of his family over Thanksgiving weekend. For as long as he could remember, his family members enjoyed making puns, mostly hackneyed and groan-inducing plays on words which nevertheless served to brighten family gatherings. So, Neely decided to recruit his family members in keeping track and rating all of the puns they produced during the weekend. Not only did he find that as the weekend progressed both the frequency and the quality of the

puns went up, but he also found that this new way of engaging with their punning made his family members feel even closer and more convivial. What he realised from this experiment was the value of group tracking, the positive effect that having a shared metric can have on people.

One of the most important aspects of the practices of self-quantification engaged in by the people I studied is the opportunity it created for them to *share* information with others. Members of the Quantified Self Movement meet up regularly to share their experiences and narratives, and many of the technologies that they use include functions that allow them to share and compare their measurements with friends and followers. The Withering Wi-Fi body scale, for example, includes a function that allows users to automatically tweet their weight and body mass index (BMI) to their followers on Twitter, and the Nike+ app allows users to share their runs in real time so their friends can send them encouragement in the form of 'cheers'. This social dimension of mediation, however, is not unique to quantified selfers. All mediation is inherently social. Whenever we appropriate a particular tool to take action, we connect ourselves to communities associated with that tool and reproduce the interaction orders and forms of social organisation that that tool helps to make possible.

The term axiology is usually used to refer to the ways we assign *value* to different things, people and happenings in our world: our theories about how the world 'ought to be' (Ladd 2015). Value, however, is ultimately a matter of the kinds of social relationships we build and the kinds of agreements we make with one another about what will be regarded as good and bad, right and wrong, normal and abnormal. When I consider the axiological dimension of mediation, then, what I am interested in is the way the technologies we use affect the kinds of societies we create and how we think we ought to treat one another. I am interested in what kinds of social beings are produced when people engage in practices like measuring, comparing, evaluating and storytelling (Abend and Fuchs 2016).

In the last two sections I talked about particular discursive processes associated with the ontological and epistemological dimensions of mediation, namely, entextualisation and semiotisation. The discursive process associated with the axiological dimension of mediation is *contextualisation*, the process through which people work together to create and negotiate the *social contexts* of their actions, and how these contexts end up affecting how they value different kinds of knowledge and different kinds of social identities. Older versions of the idea of context tended to see it as a kind of 'container' in which our interactions occur or as a function of the physical conditions of a particular situation or the rules and norms of a particular culture. Later researchers, however, especially in the fields of interactional sociolinguistics (Gumperz 1982; Tannen 1993), began to view context as more dynamic and negotiated, something that we create moment by moment in interaction. More recent work in technologically mediated communication, such as that of Tagg and her colleagues (2017), discusses the way mediational means like social media sites introduce different affordances and constraints for users to 'design' different kinds of contexts for their interactions, affordances and constraints which ultimately reflect particular ideas about what sorts of social relationships are 'normal', 'reasonable', or 'polite'. The process of contextualisation has to do with how different technologies help create and reproduce particular kinds of social relationships and 'cultural storylines' (Davies and Harré 1990) and how these social relationships and storylines come to constitute moral claims (Christians 2007).

The effect technologies have on creating social relationships and community norms can be seen not just in the examples earlier, where technologies of measuring and sharing data (about such things as moods and puns) can bring people together by giving them a shared purpose and a 'shared metric' to talk about it. It can also be seen in phenomena

like Quantified Self ‘Meetups’, where the mastery of the genre of the ‘show and tell story’ becomes the main emblem of membership in the Quantified Self community. This effect is as evident in the ‘professional sciences’ as it is in the ‘amateur’ experiments of quantified selfers. Scientists (as well as social scientists and humanities scholars) operate in ‘communities of practice’ (Lave and Wenger 1991) which share conventions about tool use, standards of measurement and ways of communicating findings. While such conventions, as I mentioned earlier, can constrain ways of experiencing and representing the world, they are also indispensable *interfaces* for social interaction between scholars (Fujimura 1992).

At the same time, technologies (and the standards they promote) also have a way of dividing and excluding people. One of the most famous examples of how technologies create (and constrain) the conditions for certain kinds of social interactions, and so impose a certain ‘moral order’ onto society, is the story of the low-hanging overpasses on the parkways of Long Island told by Langdon Winner (1980) in his famous essay, ‘Do Artifacts Have Politics?’ The underpasses, according to Winner, were designed intentionally by city planner Robert Moses to make them inhospitable to public buses, thus restricting access to Jones Beach and other leisure spots to low-income (mostly African American) residents of New York City who did not own cars. Another example is Bruno Latour’s study of the sociology of a door closer, told in the voice of a fictional engineer named Jim Johnson (Johnson 1988), in which he examines how humans and technologies work together to regulate social relationships through determining ‘what gets in and what gets out’ of particular social spaces (299). Even a device as simple as a door closer, Latour argues, is a highly moral social actor. A more recent example can be seen in the work of archivist Todd Presner (2014) on the role of algorithms in processing data for a visual history archive of Holocaust testimonies. Mediation, he argues, always has the effect of creating ‘information architectures’ which end up including and excluding not just certain kinds of information but also certain kinds of social actors. Like Latour’s door closer, algorithms, he says, tend to ‘let in some people, and not others’, resulting in a situation where ‘our analysis of the “not others” can’t be very important, certainly not central’ (n.p.).

When it comes to research, the distancing or exclusionary effects of mediation can be seen most clearly in the way technologies mediate the relationship between the researcher and the ‘subjects’ of his or her research, but it can also be seen in the stories of quantified selfers, like Miles Klee (see earlier), who found that attempts to quantify his sex life had a dramatic, and not particularly positive, effect on his relationship with his partner, and Alexandra Carmichael,⁶ who came to understand, after years of self-tracking, that her obsessive quest to understand herself was actually alienating herself from herself, making her feel like she was never quite good enough, that she always came up short of reaching her elusive goals.

All technologies, whether they be door stoppers, sex-tracking apps or text analysis programmes are ideological, and one of the main ways they exercise their ideology is by channelling their users into particular kinds of relationships with other people and by constructing for these other people certain social roles such as intruders, partners, colleagues or subjects. To use Foucault’s (1988: 18) terminology, all technologies are to some degree ‘technologies of power, which determine the conduct of individuals and submit them to certain ends’.

The exercise of power that technology makes possible includes, of course, economic power. Whenever we engage in inquiry, whether in a laboratory or with an app on our iPhones, we are participating in and helping to maintain a particular economic system. When it comes to self-trackers using commercial apps, they are not just contributing to what has come to be called ‘surveillance capitalism’ (Zuboff 2015) by making available their personal data to be sold to data brokers, advertisers, pharmaceutical companies and

other concerns, they are also helping to advance a certain approach to ‘health’ and ‘selfhood’ which emphasises individual responsibility and notions of entrepreneurial citizenship over collective responsibility and notions of social welfare (Lupton 2016; Rose 1999). In other words, through the kinds of ‘selves’ and the kinds of relationships they encourage, self-tracking apps contribute to advancing a neoliberal economic agenda.

Similarly, scholars who engage in the digital humanities, or any other academic endeavour for that matter, are wittingly or unwittingly advancing particular economic agendas designed to distribute resources and knowledge in particular ways. Much of the work in digital humanities, for example, feeds into industry-driven agendas about what constitutes valuable (‘objective’, ‘impactful’) research and what doesn’t, agendas which are reproduced by university administrators and funding authorities. The movement to ‘modernise’ the humanities through the use of digital technologies also plays into a narrative that the only way to make a humanities education beneficial to job seekers is to transform it into ‘a course of training in the advanced use of information technology’ (Allington et al. 2016: np).

Daniel Allington et al. (2016), all themselves respected scholars in the field of digital humanities, portray the ideological agenda of the field in this way:

For enthusiasts, Digital Humanities is ‘open’ and ‘collaborative’ and committed to making the ‘traditional’ humanities rethink ‘outdated’ tendencies: all Silicon Valley buzzwords that cast other approaches in a remarkably negative light, much as does the venture capitalist’s contempt for ‘incumbents.’ Yet being taken seriously as a Digital Humanities scholar seems to require that one stake one’s claim as a builder and maker, become adept at the management and visualization of data, and conceive of the study of literature as now fundamentally about the promotion of new technologies. We have presented these tendencies as signs that the Digital Humanities as social and institutional movement is a reactionary force in literary studies, pushing the discipline toward post-interpretative, non-suspicious, technocratic, conservative, managerial, lab-based practice.

Whether one agrees with this description or not, the important point is that all forms of research are also forms of political economy and that the technological tools we use to engage an inquiry are never value free; they are always somehow tied up with the distribution of economic and social capital and the sorting of social subjects into winners and losers.

Conclusion and suggestions for further research

In his essay ‘The End of Theory’, Chris Anderson, editor-in-chief of *Wired*, argues that digital technology and big data have made theory obsolete. He writes:

This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.

(2008: para 7)

The argument I have been trying to make in this chapter is that numbers do *not* ‘speak for themselves’, that they are always part of a *nexus of practice* formed by complex interactions

between human, institutional and technological actors, mediated through cultural tools (as diverse as computers, smart phones, ‘show and tell talks’ and gold stars) and dependent on the processes of *entextualisation*, *(re)semiotisation* and *(re)contextualisation* that these tools make possible. The aim of mediated discourse analysis is not to undermine the disciplinary practices, forms of knowledge or possibilities for social action that arise from technologically mediated forms of inquiry, whether they be practices of self-tracking engaged in by quantified selfers or practices of ‘distant reading’ or ‘data visualisation’ practised by scholars in the humanities and social sciences. Rather, its aim is to highlight the fact that all forms of inquiry take place within, and help to strengthen, particular *sociotechnical networks* (Star 1990: 32) which inevitably promote particular ways of knowing the world, experiencing it and valuing it, and to offer a method for untangling these networks and reflexively critiquing what we do as scholars.

What the examples and analysis I have offered demonstrate is that mediation and disciplinary paradigms are rarely as simple as they seem. Quantifying the self is always situated within a complex nexus of different kinds of practices which make use of different kinds of technologies. As Belliger and Krieger (2016: 25) point out, ‘Quantifying the self is not enough; numbers and statistics must be interpreted, that is, integrated into networks of identity, society, and meaning’. The same might be said about the digital humanities: quantification, rather than making more traditional forms of humanistic interpretation obsolete, in many ways make them even more relevant.

The methodological toolkit of mediated discourse analysis is called *nexus analysis* (Scolon and Scollon 2004). It involves attempting to understand this complex nexus of people practices and technologies through staged ethnographic research which involves (1) identifying the key actions and the key technologies which are mediating those actions in a particular social situation; (2) exploring the kinds of affordances and constraints introduced by these technologies and considering how they make certain kinds of actions more possible than others; (3) examining the ways in which mediated actions help to construct certain kinds of social identities, certain kinds of social relationships and, ultimately, certain kinds of societies; and (4) understanding how altering social actors’ awareness of their use of technology and the way it affects how they experience, know and value the world can introduce the potential to empower them to change the nexus of practice for their own benefit or for the benefit of others. All of these stages involve practices appropriated from a range of disciplines, from participant observation to ethnographic interviews, to the collection of discursive artefacts, and make use of a range of analytical paradigms from conversation analysis, to narrative analysis, to the ethnography of speaking. In this way, mediated discourse analysis itself is a nexus of practice, a kind of remix cobbled together to address the contingencies of whatever combination of practices, actors, technologies and social relationships it is presented with. The last stage – changing the nexus of practice – reveals the underlying activist agenda of mediated discourse analysis, its commitment not just to studying the world but also to changing it, and its unabashed admission that the researcher can never really position him or herself apart from the people or phenomena that he or she is researching.

What nexus analysis can contribute to our study of digital humanities is a reflective stance which compels us to consider the ways in which the tools that we use, whether they be computer programmes, information visualisations, statistical methods, analytical frameworks or the genres through which we report our results, impact the epistemological, ontological and axiological stances that we promote through our research activities. A nexus analysis can be performed as a stand-alone study in which the mediated actions of other people as they construct knowledge in the world are studied, as with the study of self-quantifiers

which I presented earlier, or it can be performed alongside nearly any other research project in which we are involved by making our own research and the means through which we conduct it the secondary object of study. In the latter case, nexus analysis can serve as a way of ‘keeping us honest’ by reminding us that whenever we do research, we are always positioned within a complex nexus of beliefs and values and technologies, and that no ‘objective’ appraisal of our findings is possible without some effort to untangle this nexus and to account for how we are positioned within it.

Advances in both digital technologies and in the theorising around digital humanities present a range of new challenges and opportunities for scholars engaged in mediated discourse analysis. New ways of using technology to capture data from ethnographic sites, for example: make available to researchers access to settings and interactions which they normally would not be able to observe. (Kjær and Larsen 2015). At the same time, these technologies add even more layers of mediation between researchers and the phenomena they are studying. Similarly, the increasing embeddedness of digital technologies into our everyday lives presents researchers with a double-edged sword, dramatically highlighting the ethical dimensions of mediation. The 2018 scandal involving Cambridge Analytica using data acquired in the context of a scholarly study to influence political campaigns, (Grasseger and Krogerus 2017; Jones forthcoming), for example, showed how scholarly institutions, funding bodies and researchers are increasingly implicated in a nexus of practice in which *value* (both economic and scholarly) is derived from extracting more and more information from people based on questionable models of consent. In her 2010 plenary to the Digital Humanities Conference in London, Melissa Terras (2010) introduced an open participatory initiative to explore the life and writings of Jeremy Bentham, inventor of the notion of the ‘panopticon’, and then used the metaphor of the panopticon to interrogate various issues arising in the digital humanities such as digital identity, professionalism and funding. But the panopticon is also in some ways an apt metaphor for the digital humanities itself, which more and more is involved the harvesting of ‘big data’ or engaging individuals in ‘crowd-sourcing’ projects like the one she described, while still struggling with formulating ethical frameworks around how to handle such data or protect the rights of individuals within the ‘crowd’.

Perhaps the key challenge for the digital humanities, and for mediated discourse analysts seeking to contribute to it, does not so much involve the digitisation of the humanities, but the digitisation of *humanity*, the increasing transformation of humans into data by algorithms and the political and social consequences of this datification (Cheney-Lippold 2017).

Both the Quantified Self Movement and the digital humanities are best thought of not as disciplines or schools of thought, but rather as metaphors for complex sociotechnical networks of people and machines, policies and institutions that produce particular ontologies, epistemologies and moral orders. And like all metaphors, they should be seen as carrying in them the power to both reveal and conceal, the power to promote affinity or alienation, conflict or co-operation (Star 1990).

Notes

- 1 <http://quantifiedself.com/2012/08/new-york-qs-showtell-17-recap/>
- 2 <http://quantifiedself.com/2012/12/amelia-greenhall-on-gold-star-experiments/>
- 3 <http://quantifiedself.com/2012/05/sky-christopherson-on-the-quantified-athlete/>
- 4 <https://vimeo.com/16691352>
- 5 <https://vimeo.com/56193255>
- 6 <http://quantifiedself.com/2010/04/why-i-stopped-tracking/>

Further reading

- 1 Jones, R. (2015). Discourse, cybernetics, and the entextualization of the self. In R. Jones, A. Chik and C. Hafner (eds.), *Discourse and digital practices: Doing discourse analysis in the digital age*. London: Routledge, pp. 28–47.
- 2 Larson, M.C. and Raudaskoski, P.L. (2018). Nexus analysis as a framework for Internet studies. In J. Hunsinger, L. Klastrup and M. Allen (eds.), *International handbook for Internet research* (Vol. 2). New York: Springer.
- 3 Molin-Juustila, T., Kinnula, M., Iivari, N., Kuure, L. and Halkola, E. (2015). Multiple voices in participatory ICT design with children. *Behaviour & Information Technology* 34(11): 1079–1091.
- 4 Scollon, R. and Scollon, S.W. (2004). *Nexus analysis: Discourse and the emerging internet*. London: Routledge.

References

- Abend, P. and Fuchs, M. (2016). Introduction: The quantified self and statistical bodies. *Digital Culture & Society* 2(1): 5–22.
- Allington, D., Brouillette, S. and Golumbia, S. (2016). Neoliberal tools (and archives): A political history of digital humanities. *LA Review of Books*, 1 May. Available at: <https://lareviewofbooks.org/article/neoliberal-tools-archives-political-history-digital-humanities/#1> (Accessed 16 November 2016).
- Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired*, 23 June. Available at: www.wired.com/2008/06/pb-theory/ (Accessed 21 January 2017).
- Bauman, R. and Briggs, C.L. (1990). Poetics and performance as critical perspectives on language and social life. *Annual Review of Anthropology* 19: 59–88.
- Belliger, A. and Krieger, D.J. (2016). From quantified to qualified self. *Digital Culture & Society* 2(1): 25–40.
- Bourdieu, P. (1992). *Language and symbolic power* (New ed.). Cambridge: Polity Press.
- Burdick, A., Drucker, J., Lunenfeld, P., Presner, T. and Schnapp, J. (2016). *Digital humanities*. Cambridge, MA: The MIT Press.
- Butler, J. (1993). *Bodies that matter*. New York: Routledge.
- Cheney-Lippold, J. (2017). *We are data: Algorithms and the making of our digital selves*. New York: NYU Press.
- Christians, C.G. (2007). Cultural continuity as an ethical imperative. *Qualitative Inquiry* 13(3): 437–444.
- Culkin, J.M. (1967). A schoolman's guide to Marshall McLuhan. *Saturday Review*: 51–53, 71–72, 18 March. Available at: <https://webpace.royalroads.ca/llefevre/wp-content/uploads/sites/258/2017/08/A-Schoolmans-Guide-to-Marshall-McLuhan-1.pdf> (Accessed 14 January 2017).
- Cayley, J. (2016). Of capta, vectoralists, reading and the Googlization of the university. In R. Simanowski (ed.), *Digital humanities and digital media*. London: Open Humanities Press, pp. 69–68.
- Davies, B. and Harré, R. (1990). Positioning: The discursive construction of selves. *Journal for the Theory of Social Behaviour* 20(1): 43–63.
- Deleuze, G. and Guattari, F. (1987). *A thousand plateaus: Capitalism and schizophrenia*. Minneapolis: University of Minnesota Press.
- Dembrosky, A. (2011). Invasion of the body hackers. *FT Magazine*, June 10. Available at: www.ft.com/content/3ccb11a0-923b-11e0-9e00-00144feab49a (Accessed 3 January 2017).
- Descartes, R. (1986). *Meditations on first philosophy: With selections from the objections and replies* (J. Cottingham, trans.). Cambridge: Cambridge University Press.
- DiClemente, C.C., Marinilli, A.S., Singh, B. and Bellino, E. (2000). The role of feedback in the process of health behavior change. *American Journal of Health Behavior* 25(3): 217–227.
- Drucker, J. (2016). At the intersection of computational methods and the traditional humanities. In R. Simanowski (ed.), *Digital humanities and digital media*. London: Open Humanities Press, pp. 41–68.

- Engel (2011). What is the difference between ‘ontology’ and ‘epistemology’? *Quora*: n.p. Available at: www.quora.com/What-is-the-difference-between-ontology-and-epistemology (Accessed 24 January 2017).
- Foucault, M. (1988). *Technologies of the self: A seminar with Michel Foucault* (L.H. Martin, H. Gutman and P.H. Hutton, eds.). Amherst: University of Massachusetts Press.
- Foucault, M. (2003). Technologies of the self. In M. Foucault and P. Raboinow (ed.), *Ethics, subjectivity and truth*. New York: The New Press.
- Frost, H. and Smith, B.K. (2003). *Visualizing health: Imagery in diabetes education*. Paper presented at the Designing for User Experiences, San Francisco, 5–7 June.
- Fujimura, J.H. (1992). Crafting science: Standardized packages, boundary objects, and ‘translation’. In A. Pickering (ed.), *Science as practice and culture*. Chicago, IL: University of Chicago Press, pp. 168–211.
- Grasseger, H. and Krogerus, M. (2017). The data that turned the world upside down. *Motherboard*, 28 January. Available at: https://motherboard.vice.com/en_us/article/mg9vvn/how-our-likes-helped-trump-win (Accessed 30 March 2018).
- Gee, J.P. (2014). *Unified discourse analysis: Language, reality, virtual worlds and video games*. London: Routledge.
- Gitelman, L. (ed.). (2013). *‘Raw data’ Is an oxymoron*. Cambridge, MA: The MIT Press.
- Gibson, J.J. (1986). *The ecological approach to visual perception*. Boston: Psychology Press.
- Gumperz, J.J. (1982). *Discourse strategies*. Cambridge: Cambridge University Press.
- Halpern, O. (2014). *Beautiful data: A history of vision and reason since 1945*. Durham, NC: Duke University Press.
- Hayles, N.K. (2012). *How we think: Digital media and contemporary technogenesis*. Chicago and London: University of Chicago Press.
- Hume, D.E. (1985). *A treatise of human nature*. London: Penguin Books.
- Iedema, R. (2001). Resemiotization. *Semiotica* 137(1–4): 23–39.
- Irigaray, L. (1985). *Speculum of the other woman*. Ithaca: Cornell University Press.
- Johnson, J. (a.k.a. Bruno Latour) (1988). Mixing human and non-humans together: The sociology of a door closer. *Social Problems* 35(3): 298–310.
- Jones, R. (2007). Good sex and bad karma: Discourse and the historical body. In V.K. Bhatia, J. Flowerdew and R. Jones (eds.), *Advances in discourse studies*. London: Routledge, pp. 245–257.
- Jones, R.H. (2009). Dancing, skating and sex: Action and text in the digital age. *Journal of Applied Linguistics* 6(3): 283–302.
- Jones, R. (2010). Cyberspace and physical space: Attention structures in computer mediated communication. In A. Jaworski and C. Thurlow (eds.), *Semiotic landscapes: Text, space and globalization*. London: Continuum, pp. 151–167.
- Jones, R.H. (2013a). *Health and risk communication: An applied linguistic perspective*. London: Routledge.
- Jones, R. (2013b). *Cybernetics, discourse analysis and the entextualization of the human*. A plenary address delivered at the 31st AESLA (Spanish Applied Linguistics Association) International Conference, La Laguna/Tenerife, Spain, 18–20 April.
- Jones, R. (2013c). *Digital technologies and embodied literacies*. A Paper presented at the 5th International Roundtable on Discourse Analysis, Hong Kong, 23–25 May.
- Jones, R. (2015a). Discourse, cybernetics, and the entextualization of the self. In R. Jones, A. Chik and C. Hafner (eds.), *Discourse and digital practices: Doing discourse analysis in the digital age*. London: Routledge, pp. 28–47.
- Jones, R. (2015b). *The rise of the datasexual*. A Paper presented at Language and Media 6, Hamburg, Germany, 7–9 September.
- Jones, R. (2016). Mediated self-care and the question of agency. *Journal of Applied Linguistics and Professional Practice* 13(1–3): 167–188.
- Jones, R. (forthcoming). The rise of the Pragmatic Web: Implications for rethinking meaning and interaction. In C. Tagg and M. Evans (ed.), *Historicising the digital*. Berlin: De Gruyter Mouton.

- Jones, R.H. and Hafner, C.A. (2012). *Understanding digital literacies: A practical introduction*. Milton Park, Abingdon, Oxon and New York: Routledge.
- Kitchin, R. and Dodge, M. (2011). *Code/Space: Software and everyday life*. Cambridge, MA: The MIT Press.
- Kjær, M. and Larsen, M.C. (2015). *Digital methods for mediated discourse analysis*. A Paper delivered at ADDA 1: 1st International Conference: Approaches to Digital Discourse Analysis. Valencia, Spain, 19–20 November.
- Klee, M. (2014). I tried to quantify my sex life- and I'm appalled. *The Daily Dot*, 8 February. Available at: <https://kernelmag.dailydot.com/issue-sections/headline-story/11626/quantify-sex-app-love-tracker/>
- Kress, G. (2009). *Multimodality: A social semiotic approach to contemporary communication* (1st ed.). London and New York: Routledge.
- Ladd, G.T. (2015). *Outlines of metaphysic: Dictated portions of the lectures of Hermann Lotze*. South Yarra, Australia: Leopold Classic Library.
- Latour, B. (1999). *Pandora's hope: Essays on the reality of science studies*. Cambridge, MA: Harvard University Press.
- Latour, B. (2009). Tarde's idea of quantification. In M. Candea (ed.), *The social after Gabriel Tarde: Debates and assessments*. London: Routledge, pp. 145–162.
- Lave, J. and Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.
- Lemke, J.L. (1999). Typological and topological meaning in diagnostic discourse. *Discourse Processes* 27(2): 173–185.
- Lupton, D. (2016). *The quantified self*. Cambridge, UK: Polity Press.
- Manovich, L. (2013). Museum without walls, art without names. In C. Vernallis, A. Herzog and J. Richardson (eds.), *The Oxford handbook of sound and image in digital media*. Oxford: Oxford University Press, pp. 253–278.
- Moretti, F. (2013). *Distant reading*. London and New York: Verso Books.
- Norris, S. and Jones, R.H. (2005). *Discourse in action: Introducing mediated discourse analysis*. London: Routledge.
- Nishida, K. (1958). *Intelligibility and philosophy of nothingness*. Tokyo: Maruzen.
- Presner, T. (2014). The ethics of the algorithm. *Institute news*. USC Shoah Foundation Institute for Visual History and Education. Available at: <https://sfi.usc.edu/news/2014/06/5471-ethics-algorithm> (Accessed 14 January 2017).
- Rose, N. (1999). *Powers of freedom: Reframing political thought*. Cambridge: Cambridge University Press.
- Rosenbloom, P.S. (2012). Towards a conceptual framework for the digital humanitie. *Digital Humanities Quarterly* 6(2).
- Scollon, R. (2001). *Mediated discourse: The nexus of practice*. London: Routledge.
- Scollon, R. and Scollon, S.W. (2004). *Nexus analysis: Discourse and the emerging internet*. London: Routledge.
- Simanowski, R. (2016). Introduction. In R. Simanowski (ed.), *Digital humanities and digital media*. London: Open Humanities Press.
- Spinoza, B. (1996). *Ethics* (E. Curley, trans.). London: Penguin Classics.
- Star, S.L. (1990). Power, technology and the phenomenology of conventions: On being allergic to onions. *The Sociological Review* 38(S1): 25–56.
- Tagg, C., Seargeant, P. and Brown, A.A. (2017). *Taking offense on social media: Conviviality and communication on Facebook*. Basingstoke: Palgrave Macmillan.
- Tannen, D. (1993). *Framing in discourse*. New York: Oxford University Press.
- Terras, M. (2010). *Present, not voting: Digital humanities in the panopticon*. A Plenary address delivered at the Digital Humanities 2010 Conference, Kings College, London.
- Vygotsky, L.S. (1962). *Thought and language*. Cambridge, MA: MIT Press.
- Winner, L. (1980). Do artifacts have politics? *Daedalus* 109(1): 121–136.

- Wolf, G. (2010). The data driven life. *The New York Times Magazine*, May 2. Available at: www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html?pagewanted=all (Accessed 12 January 2017).
- Wouters, P., Beaulieu, A., Scharnhorst, A. and Wyatt, S. (2012). *Virtual knowledge: Experimenting in the humanities and the social sciences*. Cambridge, MA: The MIT Press.
- Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology* 30(1): 75–89.

Critical discourse analysis

Paul Baker and Mark McGlashan

Introduction

Critical discourse analysis (CDA) is a methodological approach to the analysis of language in order to examine social problems, with a focus on power, particularly issues around abuses of power, including discrimination and disempowerment. Analysis considers the language in texts but also considers the texts in relationship to the wider social context in which they are produced and received. Many critical discourse analysts carry out close readings of a small number of texts, focusing on how linguistic phenomena like metaphor, agency, backgrounding, hedging and evaluation can help to represent various phenomena or present arguments for different positions. This chapter focusses on how automatic forms of computational analysis (in this case, corpus linguistics) could be used in combination with the kind of ‘close reading’ qualitative analyses associated with CDA, particularly in cases where analysts are working with large databases containing hundreds or thousands of texts.

The type of database that this chapter focusses on is online news. News is particularly relevant for studies in CDA, as it tends to involve reports of social issues, and the way that news is reported can have important consequences for societies. Van Dijk (1991) argues that newspapers influence public opinion, whereas Gerbner et al. (1986) contend that the media’s effect on an audience compounds over time due to the repetition of images and concepts. Searchable online databases of news articles like Lexis Nexis and Factiva have enabled researchers within the digital humanities to quickly gain access to vast datasets of stories around particular topics. Users can conduct targeted searches for articles from certain newspapers or time periods, as well as specify words that articles must contain. Such databases have been used effectively in critical-oriented research, for example, on gay men (Baker 2005), refugees and asylum seekers (Baker et al. 2008) and Muslims (Baker et al. 2013). By employing software to identify linguistic frequencies in corpora (essentially large representative collections of electronically encoded texts), it is possible to make sense of data comprising millions of words. This enables generalisations to be made of the whole, as well as directing researchers to typical (or atypical) cases so that more qualitative, close readings can be carried out. The three studies mentioned earlier just focused on the articles themselves, although more recently, McEnery et al. (2015) examined both news articles and

discussion on the social media platform Twitter relating to the ideologically inspired murder of the British soldier Lee Rigby.

In the last two decades the way that people engage with print news has changed; as (access to) technology has changed, so, too, have practices. People are now able to access information (including news) more readily on mobile and Internet-connected devices, which has ‘shifted the roles traditionally played by newspapers, television stations, [etc.]’ (Westlund and Färdigh 2015) and online news now offers an alternative to print (Thurman 2018). With these new technologies, breaking news can be almost instantaneously reported and, apart from a small number of news outlets which charge subscriptions for access to their content, can be accessed largely free of charge. Newspaper editors are able to gather large amounts of data about their readers in terms of which links they click on and how long they remain on a web page (as well as other forms of ‘customer analytics’). Such information can then be used in order to refine news content to prioritise the sorts of stories that readers are most likely to read. However, perhaps the most marked change in news production and consumption involves the fact that readers can publish their own reactions to stories, as well as interact with others, creating dialogue between the news producer and other readers. As most user comments appear after the article itself, the phrase ‘below the line’ is sometimes used as a shorthand to refer to this user-generated discussion.

In line with work by Chovanec (2014), we argue that reader comments have the potential to alter the way that news articles are engaged with and understood. They offer readers the opportunity to engage (passively or actively) within a community of practice (cf. Wenger 1998) which involves other readers, and to learn about different types of responses to news stories. It cannot be argued that reader comments represent the most typical views of the public or even that they are typical of those who read the article as people who take the time to comment on articles may have stronger opinions than most, or may have other motives (such as trolling other readers). However, we argue that the comments have the potential to influence opinion, particularly if a large number of comments put forward the same argument across multiple articles on the same topic.

Therefore, an aim of this study is to acknowledge that digital engagement with news involves an additional dimension, one that is important to consider when carrying out analyses of the news articles themselves, and one which the digital humanities is well placed to examine – user-generated comments.

In this chapter we first briefly describe critical discourse analysis, its aims and methods and how it can be utilised on large collections of digital texts or corpora. We then describe the focus of our research, the analysis of newspaper articles and reader comments taken from the British newspaper the *Daily Express* which refer to Romanians. After outlining why this topic is of relevance to critical discourse analysts and providing information relating to the social context of the topic, we list our research questions and discuss how the corpus was collected.

This is followed by an illustrative analysis of the corpus data, applying the keywords technique to identify salient words in the articles and comments, and then using a second technique, collocation to gain an idea about how particular keywords are used in context. To illustrate how our method works we focus on a small number of keywords (*Romanians*, *them* and *us*), identifying collocates of these words which help to give an impression of how they are used in context and contribute towards ideological positions. We then show how the tool ProtAnt can be used in order to rank the news articles in order of prototypicality, helping us to focus on a ‘down-sampled’ set of articles which are the most representative of the entire set. We are thus able to carry out a close qualitative analysis of the language in a

single prototypical article and then compare its content against the reactions of commenters. The chapter concludes with a discussion of potential issues surrounding taking a critical approach within the digital humanities, and consideration of future directions for this field.

Background

CDA is an ‘interdisciplinary research movement’ interested in examining ‘the semiotic dimensions of power, injustice, abuse, and political-economic or cultural change in society’ (Fairclough et al. 2011: 357). Wodak and Meyer (2009: 10) define CDA as aiming ‘to investigate critically social inequality as it is expressed, constituted, legitimized, and so on, by language use (or in discourse)’. Developing from the ‘East Anglia School’ of critical linguistics (Fowler et al. 1979), which involved analysis of texts in order to identify how phenomena are linguistically represented, and strongly influenced by the Frankfurt School, particularly the critical work of Habermas (1988), CDA sees discourse as semiotic and social practice (Fairclough 1992, 1995, 2003), constituting both modes of social action and representation. Where CDA differs from critical linguistics is that analysis goes beyond the text to consider various levels of context, including the processes of production and reception around the text, the social, political, economic and historic context in which the text occurred, and the extent to which texts contain traces of other texts through direct quotation, parody or allusion (intertextuality, cf. Kristeva 1986) or involve interdiscursivity (e.g. the presence of discourse associated with advertising might be found in an educational text, cf. Fairclough 1995).

A number of different ‘schools’ or approaches to CDA exist, each with different (and sometimes overlapping) theoretical and analytical foci and toolsets, for example, Fairclough’s (1992, 1995, 2003) dialectical-relational approach, Reisigl and Wodak’s (2001) discourse-historical analysis, Van Dijk’s (2006) sociocognitive approach and van Leeuwen’s (1996, 1997) focus on social actor representation. Within these schools there is no single step-by-step guide to analysis, but research questions and datasets can influence the analytical procedures which are selected and the order in which they are carried out.

A relatively new form of CDA – corpus-assisted critical discourse analysis – was conceived as a way of handling large collections of texts. This approach, which draws directly on theory and methods from corpus linguistics (CL), can be seen as being directly related to the field of digital humanities. Jensen (2014) argues that ‘corpora are *per se* digital [and that] building a corpus is inherently a DH [digital humanities] project’ but that ‘CL is on the fringes of contemporary DH, which is itself currently on the fringes of the humanities’. The exploration of this method here therefore attempts to position corpus-assisted critical discourse analysis as a relevant (although potentially fringe) approach to DH scholarship. Moreover, corpus-assisted critical discourse analysis enables analysts to draw their observations from a wide range of texts rather than ‘cherry-picking’ a few texts which prove a pre-conceived point and thus attempts to mitigate researcher bias (Widdowson 2004). Important to this approach are corpus techniques based on frequency, statistical tests and automatic annotation (e.g. of grammatical categories), which enable frequent linguistic patterns to be identified within texts while concordance tools allow a qualitative reading to be carried out across multiple occurrences of a particular pattern.

The model proposed by Baker et al. (2008) advocates moving forwards and backwards between corpus techniques and close reading in order to form and test new hypotheses. The analysis is supplemented with background reading and the sorts of contextual interrogation typical of other forms of CDA. In dealing with large sets (which we define as any set that

is beyond the capabilities of the researcher or research team to carry out a close reading of every text) of electronically encoded texts, this approach is particularly germane to the digital humanities and in the analysis which follows we take a corpus-assisted approach to both identify large-scale patterns in corpora of newspaper articles and readers' comments, and to enable us to focus on a smaller number of prototypical articles for a more detailed close reading. In the following section we outline the topic under investigation and explain why it is relevant to take a CDA approach.

Background and data collection

On 23 June 2016, a national referendum was held across the UK to decide whether the country should 'Remain' in the political and economic union of member states that are located mainly in Europe, known as the European Union, popularly referred to as Brexit (British Exit). Over 33 million votes were cast, with the 'Leave' vote winning by a narrow margin (51.89 percent to 48.11 percent). While the split decision indicated a general feeling of disunity, the reasons given for people wanting to leave suggested more specific social problems. For example, in *The Guardian*,¹ Mariana Mazzucato argued that people felt aggrieved by the government's programme of austerity to reduce borrowing since the global economic crash in 2008. On the other hand, the *Daily Mail*² reported that Brexit was blamed on the German chancellor Angel Merkel's 'open-door approach' to immigration. A poll of over 12,000 people carried out by Lord Ashcroft (a former deputy chairman of the Conservative Party who now works as a major independent public pollster of British public opinion), after the vote gave the top three reasons for voting Leave as 'decisions about the UK should be taken in the UK', 'the best chance for the UK to gain control over immigration and its borders' and 'remaining meant little or no choice about how the EU expanded its membership or powers'.³

Clearly, concerns about immigration were an important factor in the decision, and in this indicative study we investigate how attitudes towards immigration may have been influenced by the ways that the media constructed immigration in the years leading up to the Brexit vote. Prior to this research, a large corpus-assisted CDA research project focused on the words *refugee*, *asylum seeker*, *immigration* and *migrant* (and their plurals) in the national British press between the years 1996 and 2005 (see Baker et al. 2008, Gabrielatos and Baker 2008). Our study takes a different approach by examining the representation of a single national identity group, Romanians. This group was chosen because Romania was a relatively recent entrant to the European Union, joining in 2007. Romania's entrance into the EU was not straightforward, indicating that there were concerns about its membership in some quarters. For example, the EU had imposed monitoring of reforms based on the rule of law, judicial reform and the fight against corruption, as well as putting in place a transitional cap on migration meaning that Romanians were not able to become resident in the UK until the beginning of 2014 (unless they were self-employed or worked in seasonal jobs). The change to the law resulted in headlines such as 'Sold Out! Flights and buses full as Romanians head for the UK',⁴ a story which was later found to be untrue and amended.⁵ These existing indicators of concern around Romania's entry to the EU make it a particularly relevant country for examination of media discourse. However, unlike most previous corpus-based work on press discourse, this study also considers readers' comments on the online versions of articles. The online comments act as type of proxy for reader reception and thus fit well with CDA's remit for considering different forms of context around texts.

Our research questions were:

- 1 How were Romanians typically linguistically represented by the *Daily Express* in the period up to Brexit?
- 2 What were the differences/similarities in linguistic representation between articles and reader comments?

To answer these questions, we carried out a corpus-assisted CDA of the corpora of articles and comments, as well as conducting a qualitative analysis of the articles and comments identified as most representative of the whole.

As this is a relatively small-scale study aimed more at demonstrating principles and techniques of analysis, we have chosen to analyse just one newspaper, *The Daily Express*, a right-leaning or conservative tabloid.⁶ We collected data from approximately a decade-long period (24 July 2006 until 23 June 2016, which was the day of the Brexit vote). The articles and their corresponding comments that form the data were identified using the search functionality of the *Daily Express* website (www.express.co.uk/search) which allows users to search all online pages that contain a particular search term and gives several methods for refining and ordering searches by date, time, section and headline. The search used here returned a list of 1945 articles published online during ‘All Time’ and containing the search term ‘romanian’. This list displays a headline, sub-headline, image, paper section and time/date of publication referring to each page, which, when clicked, uses a hyperlink to navigate to the appropriate page. It would be possible to manually click through the entire list and collate, by copying and pasting, all of the articles and comments sections for each page. However, this process would be arduous and open to human error. Instead, computational methods were implemented to harvest data automatically from these pages – an activity typically referred to as ‘web scraping’ – using the Python⁷ programming language and the Python libraries Beautiful Soup⁸ and Selenium.⁹ Beautiful Soup enables users to parse and scrape data from the Hypertext Markup Language (HTML) code from which web pages are made and Selenium gives the ability to simulate user behaviour (e.g. scrolling and clicking links) and automate activity in web browsers such as Mozilla Firefox¹⁰ and Google Chrome.¹¹ Concerning ethics, this approach simply emulates behaviours that could be carried out manually and only targeted data published for free, public, online consumption. No data collected were shared or redistributed in any way, and no personally identifiable information (e.g. comment/article author names/handles) were targeted for collection. The web scraping programme developed for this data collection activity first identified the hyperlink for each of the 1945 pages returned, assigned a number to each page and then navigated to each page in turn. The programme then extracted the HTML code for the page that it navigated to and identified and saved the headline, sub-headline and the body text for the article on that page into a text file. If an article had any comments, they were saved into another text file. The files for articles and comments were assigned the same number as the page from which they were extracted and saved into separate ‘articles corpus’ and ‘comments corpus’ folders. The process is outlined graphically in Figure 13.1.

As data was collected during July 2016, articles published after the sampling cut-off date of 23 June 2016 were also returned. One thousand nine hundred and twenty-nine articles remained in the sample following the removal of those articles published after the cut-off date from which an articles corpus (771,878 words) and comments corpus (2,166,148 words) remained.¹²

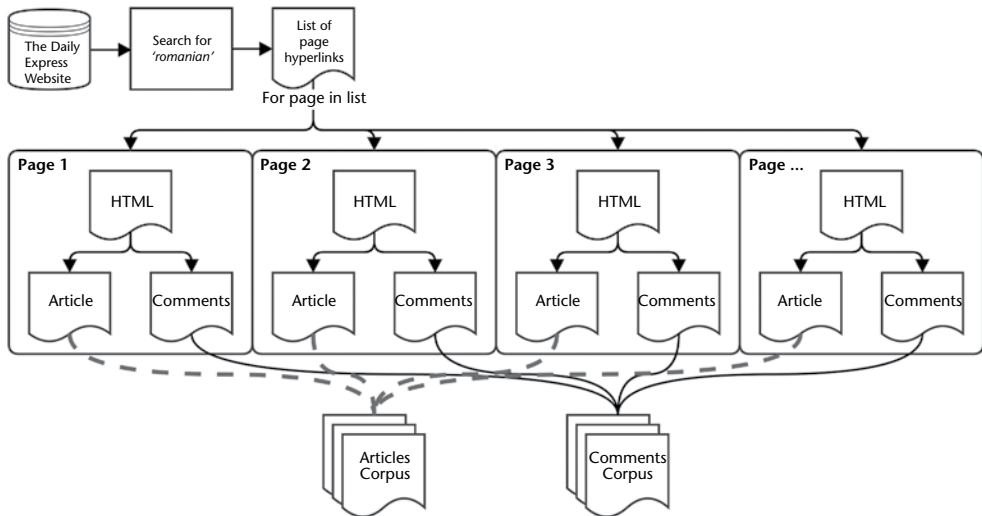


Figure 13.1 Process for extracting articles and comments corpora

Programming languages like Python and web scraping technologies provide important automatable methods for researchers in the digital humanities. They enable huge amounts of data, such as the corpora described earlier, to be collected relatively quickly. However, use of these methods requires an acknowledgement of several issues. *Vis-à-vis* time, learning how to use programming languages (even for *ad hoc* purposes) is a lengthy process, and the data collection itself is not instantaneous – it can take several hours to collect corpora such as those used in this study. Time is also a factor when considering data ‘cleaning’. The data collected for this study, for example, required manual investigation to check that the programme collected what was required. In the case of the user comments there was an issue where the first few words in a comment were repeated and followed by “\r\r” (e.g. ‘Strange how the police will act when the prob\r\rStrange how the police will act when the problem affect the rich and powerful!’). This error was likely due to a redesign of the way in which comments were displayed on *The Express* website but the repetition of words would have adversely impacted on how word frequencies were calculated which is fundamental to CL studies. An additional script had to be developed to automatically identify and delete these instances of repetition.

Another issue concerned annotation. As the contents of files in the articles and comments corpora contained different kinds of data, different annotation schemes had to be developed to delineate different parts of a document. For example, in the articles corpus, headlines were placed between tags similar to those found in XML documents, for example, “<Headline> EU workers rushing to get jobs in Britain ahead of potential Brexit </Headline>”; sub-headlines and the body text of an article in these files were annotated similarly. In the comments corpus, a similar annotation scheme was used to simply distinguish one comment from another in the order in which comments were posted on the article from most to least recent. However, some comments were nested as replies to other comments, creating a conversation hierarchy. One issue with the annotation scheme used was that this hierarchy was not preserved, thus potentially decontextualising some comments from their original conversational structure.

Analysis

First, the tool AntConc version 3.4.4 (Anthony 2016) was used in order to obtain frequency lists of the datasets and obtain keywords. This tool has been chosen for analysis because it is free and relatively easy for beginners to learn how to use. A potential disadvantage is that it does not offer as wide a range of processes or statistical tests as other tools like WordSmith Tools,¹³ #LancsBox,¹⁴ CQPweb¹⁵ or Sketch Engine¹⁶ but for the purposes of introducing new analysts to the field, it is useful. Keywords are words which appear in one corpus statistically significantly more often than another, and are useful in identifying repeated concepts which may be overlooked by human analysts. They can help to focus analysis on a small set of salient words which are then subjected to a more detailed qualitative interrogation. Using the keywords facility of AntConc, comparing a word list of the articles against the comments (and vice versa) produces words which are more frequent in one corpus as opposed to another, which is useful for identifying the main lexical differences between two corpora, but will overlook similarities. To focus on both difference and similarity, both corpora were separately compared to a third dataset, a reference corpus of general written published British English (the 1-million-word BE06, see Baker 2009). The BE06 corpus offers a reasonably good reference for the way that written British English was used around the period that the news articles and comments were written. A potential issue is that it does not include informal computer-mediated communication (CMC) so comparing it against comments could result in keywords appearing that are more concerned with the register of CMC (such as the acronym *lol*) as opposed to the topic of immigration. However, this concern proved to be unfounded, and none of the keywords obtained from the comments corpus were due to the aspects of the CMC register.

In AntConc keywords are calculated by performing log-likelihood tests, taking into account the relative frequencies of all the words in the two corpora being compared as well as the sizes of each corpus.¹⁷ Table 13.1 shows the 50 keywords which have the highest log-likelihood scores (in other words, these are the most ‘key’ words) for both the articles and comments.¹⁸

For the top 50 keywords in each corpus, 19 are shared, while each corpus has 31 unique keywords. While it would have been possible to categorise the keywords according to say, an existing automatic semantic tagging system like USAS, this would have focused on dictionary definition meanings of words. A more nuanced categorisation scheme, based on the context in which the words were used, was more helpful in terms of identifying the themes that the articles accessed (see Baker 2004). For example, the keyword *David* could have been classified as a proper noun or a male name using an automatic categorisation scheme. However, for the purposes of our analysis it is more useful to know that *David* almost always refers to the former Prime Minister of the UK David Cameron and thus indicates discussion of Romanians in a political context. Therefore, the set of categories was created by subjecting the keywords to concordance analyses (involving reading tables containing citations of every keyword – see Table 13.2 for an example of a sample of some of the mentions of the word *influx* of the total that were examined) and then categorised into groups based on theme. This is an admittedly subjective approach, and in such circumstances it is useful to have more than one analyst (which is the case here) to work on the scheme.

This resulted in four main themes: Politics, Nationality, Movement and Football. In most cases categorisation into these themes was easy as words were used unambiguously (as Table 13.2 suggests, *influx* normally referred to movement of people). One word which was slightly more difficult to classify was *Manchester*. Concordance analysis revealed that 95 percent of cases referred to football teams (Manchester City and Manchester United).

Table 13.1 Keywords in articles and comments

Categories	Articles	Comments	Both
Politics	David	government, Nigel, politicians, vote	UKIP, Cameron, Labour, Farage
Nationality	Eastern, European, Bulgaria, Bulgarian, Bulgarians, Roma, Bucharest, country	country	EU, Romania, Romanian, Romanians, Britain, British, UK, Europe, countries
Movement of people	influx, migration, migrant		migrants, immigration, immigrants
Football	champions, Chelsea, Manchester, game, win, league		
Grammatical words	after, who, against	they, we, our, all, you, this, them, us, here, why, what, if, no, not	
Other adjectives	false, last	just	
Other verbs	said, has	come, want, get, should, have, do, are, is, stop, don't	will
Other nouns	workers, police, express, year, restrictions, million	people	benefits, racist

Table 13.2 Concordance table of ten citations of the keyword *influx* in the Articles corpus

Highlighting the impact of such an the whole truth about the annual find it harder to cope with an when their neighbourhood sees a large people whose neighbourhoods see an workers into the town. The huge came into the country during the recent benefit pledge and reduce the foreign "Others voiced concerns that the huge desperate bid to control huge migrant	influx influx influx influx influx influx influx influx influx influx	on the NHS, Vote Leave forecast A& of newcomers. Alp Mehmet, vice chair of children who do not speak English. of migrants. While that might seem of migrants tend to be made unhappy began in 2004 when Britain opened its from Syria of being the culprits. Town . Denying that the Prime Minister has of people would put additional strains . TANKS and riot police have been

However, a small number of cases referred to the city itself (e.g. Car cleaner Adrian Oprea began selling the Big Issue when he came to Manchester from Romania in 2009 with his wife and son). As such cases were a clear minority, the word *Manchester* was categorised as referring to the theme of football.

Not all words could be clearly categorised into themes, so a category of closed-class 'grammatical words', consisting of pronouns, determiners, prepositions, wh-questions and negators was also created. This left a smaller set of adjectives, verbs and nouns which appear at the bottom of the Table 13.1. Some of these words are suggestive of themes too (e.g. *police*

could be put in a new category of ‘Crime’, or *workers* and *benefits* in ‘Economics’) but this is a preliminary stage of analysis and was carried out in order to understand the themes underlying reference to Romanians which are indicated by multiple words referencing the same concept. The categorisation also helps us to focus on similarities and differences between the two corpora. For example, it is notable that the words about football are only keywords in the news articles, not the comments, indicating that commentators do not seem to be as interested in responding to articles which refer to Romanians in the context of football.

When *The Express* does refer to Romanians in the context of football, it is usually to mention a team or player, for example,

Romanian side CFR Cluj-Napoca, in Group F along with Manchester United, notched a 2–0 win away to Braga

(Article 1191)

The articles about football do not discuss Romanians in the context of immigration and for that reason will not play a large role in our analysis. However, it is worthwhile discovering their existence because they indicate that *The Express* did not always discuss Romania in the context of immigration, and also that these articles do not tend to provoke much reader comment.

A full keyword analysis would involve taking the words in Table 13.1 in turn and subjecting each to a more detailed analysis of context. This means examining a word’s collocates (words which frequently occur near or next to one another, usually more than would be expected by chance), as well as concordance lines, in order to obtain an impression of what a word is being used to achieve, particularly in relation to the research questions asked. For illustrative purposes, we have selected a smaller number of words from Table 13.1 for this kind of analysis. These words are *Romanians*, *us* and *them*.

Romanians

It is unsurprising that *Romanians* is a keyword – it actually appears as the focus of Research Question 1 and was also one of the search words used in creating the corpus. Due to space restrictions, we focus our analysis on this word, although note that it would be useful to consider related word forms like *Romania* and *Romanian*. As a plural noun, *Romanians* is an example of collectivisation, a type representation that involves assimilation of individuals into a group (see van Leeuwen 1996). Ideologically, collectivisation can be relevant to focus on as it can result in differences between individuals being obscured and generalisations being made. The word *Romanians* occurs 890 times in the articles and 1433 times in the comments. It is notable that in the comments *Romanians* is more frequent than *Romania* and *Romanian*, while in the articles it is less common than those words, indicating a potential difference between journalists and readers – the readers appear more likely to engage in this form of collectivisation. As a way of obtaining an impression of how *Romanians* is used in context, we can examine its collocates in both corpora. We calculated collocates using the mutual information (MI) statistic (which shows the strength of a relationship between two words). We considered all words five words on either side of *Romanians* which had an MI score of 6 or above (following Durrant and Doherty (2010: 145), who had carried out experiments on human subjects and indicated that this was a cut-off where collocates became ‘psychologically real’). As the MI measure can produce collocates that are of low

frequency (and are thus not very representative of the corpus as a whole), we stipulated that a pair must occur at least five times together before we could consider it as a collocate.

As with the keywords, collocates of *Romanians* have been grouped into thematic categories. While not an analysis in itself, this helps the analyst to spot repeated associations and also differences and similarities between the two corpora. For example, it is notable that both articles and comments tend to quantify Romanians and focus on their movement, for example,

The most reliable estimates have always come from the Migration Watch think tank. It **estimates** that about 50,000 **Romanians** and Bulgarians will arrive on our shores every year for the first five years when we lift border restrictions. If they are correct that will mean 250,000 Romanians and Bulgarians arriving.

(Article 981)

We are just over 4 months away from “Entry Day” when potentially tens of thousands, sorry I meant **hundreds** of thousands of **Romanians** & Bulgarians can legally enter Britain and the three main party leaders haven’t a clue what to do.

(Comments, 854)

The comment above was noteworthy because it appeared eight times in the comments corpus, all in the same comment thread. It appears that this comment had been made repeatedly by the same person. This was an unexpected feature of the comments data, where we found that some collocates were the result of repeated comments. Not all cases involved repetition to the same thread. For example, in Table 13.3 the collocates *errr* and *um* with *Romanians* were the result of the following comment:

Interview goes along these lines . . . Hello “ANY” MPs. Are these people Romanians or Roma gypsy’s? . . . errr well um . . . errr we welcome Romanians & Bulgarians & i’m off to the airport to have a photo shoot.

This comment appeared six times across four different threads, made by the same commenter, suggesting that it had been cut and then pasted multiple times. Analysis of clusters (e.g. fixed sequences of words occurring multiple times) revealed a small number (18) cases across the corpus where the same comment appeared at least five times. Such cases ultimately do not detract from the overall findings of the analysis, and where collocates could be attributed to such cases, they should be noted, but they do indicate a way that individuals can take advantage of the affordances of digital data by ensuring that a particular message they want to convey reaches a wider audience. Obviously, if we encounter such cases, we may want to indicate that a repeated pattern is not necessarily widely shared among a discourse community but more due to a single commenter with a particular axe to grind.

Some collocates suggest a different sort of focus between the news reports and comments. Consider *steal* and *stealing*, which are collocates of *Romanians* in the comments corpus. Collectively, they occur 15 times with *Romanians* (15/1433; a relative frequency of 0.01). A detailed analysis of concordance lines revealed five cases that refuted the idea that Romanians steal, for example, ‘There are many honest Romanians who work and do not steal’, while the other ten constructed Romanians as stealing. Even within this distinction, there were differences, with Romanians accused of stealing jobs, EU funds, UK benefits and in one case, a lawnmower from someone’s garden.

Table 13.3 Collocates of *Romanians*

Category	Articles	Comments	Both
Quantification	percent, estimates, estimate, eight, number, tens, nearly	hundreds	
Movement	arriving, settle, gain, arrived	enter, moved, arrive	
Nationalities and groups	Lithuanians, group	Latvians, Indians, Indian, Bulgarian, Roma, gipsies, gypsies, gypsy, Africans	Bulgarians, Poles
Law and Order	visa, loophole, attacks, gang, curbs, restrictions, controls	legally, steal, stealing, honest, decent, metropolitan	arrested
Homelessness	rough, homeless	begging	sleeping
Time	night, minutes, yesterday, temporary		January
Grammatical words		between	
Images		picture	
Reporting Verbs	revealed, showed, warned	showing	
Misc Verbs	lifted, expected, seeking, applied, imposed, gain	meant, met	
Others	marble, arch, league, abroad	errr, um, neighbours, difference, educated, guys	

One of the comments which challenges the idea of stealing Romanians purports to be from someone who identifies as Romanian:

And by the way you got so much youth unemployment because you're lazy and have so big requirements that nobody wants to work with you. Show me a brit girl who would wipe your elders **** when they can't do it by themselves. Show me a brit mopping the streets of your cities, show me a brit doing all these kinds of low esteem jobs and then come tell me about **ROMANIANS stealing** your **** jobs.

(747)

The analysis indicates that a collocational pair like *Romanians* + *stealing* does not indicate a straightforward 'discourse position' but instead reveals a range of potential stances. This is important to note, although we should also indicate that the more popular stance is one of Romanians as thieves, and this is something that commenters have to orient towards. On the other hand, the articles do not directly refer to Romanians stealing by using these collocates (there is not even one case where *steal* or *stealing* appears near the word *Romanians* in the articles corpus). With that said, the words *steal* and *stealing* do occur within the news articles (in positions other than near the word *Romanians*), a total of 122 times (an average of 0.06 times per article), and they do typically refer to people from Romania involved in crime.

Additionally, in the articles, other law and order collocates exist like *gang*, *curbs*, *restrictions* and *controls*. The latter three words involve citations of *The Express*'s stance that there

should be restrictions on Romanians arriving or working in the UK. This often involves quoting from others who are in favour of such restrictions.

However, critics fear a fresh wave of migration next year when visa **restrictions** on millions of **Romanians** and Bulgarians are lifted.

(Article 1037)

The efficiency with which the police destroyed huts and tents – using a digger to crush caravans – led some observers to question why British authorities can't be as ruthless. Britain, like France, has transitional **controls** on **Romanians** seeking to settle in the UK. Until next year, only Romanian migrants who have a job, or can support themselves, are allowed to stay in Britain.

(Article 1202)

Warning over 'flood' of immigrants. . . . Temporary **curbs** were imposed on **Romanians** and Bulgarians in 2005 to protect the British labour market.

(Article 1109)

Gang clearly contributes towards the discourse of Romanians as criminals, with seven of nine instances referring to crime, for example, 'A gang of Romanians has been jailed for a total of 24 years after a string of crimes across the UK' (Article 1181). The other two cases refer to *gang* negatively so could more indirectly contribute towards an association of Romanians with crime. Article 136 has the headline 'POLICE raiding a suspected slavery gang found 40 Romanians crammed inside a three-bedroom house'. Reading the article, the Romanians are constructed as victims of the gang, whose nationality is not specified. The ninth instance involves a case of a couple whose home was taken over by squatters. It contains the line 'When Samantha asked the police if the gang were Romanians on benefits she was ticked off for being "racist"'. The article refers to Romanian squatters elsewhere and is sympathetic towards Samantha, putting *racist* in distancing quotes and implying that the police were not doing their jobs properly.

Not all the Law and Order collocates of *Romanians* are negative. Consider *honest* and *decent* which are collocates in the comments corpus. These cases involve descriptions of Romanians as *honest* (11 cases) and *decent* (12 cases). However, closer reading indicates that these cases tend to make a distinction between Romanians and gypsies. The following example (which also appears to be written by someone who identifies as Romanian), is a response to news article 386 which describes a television programme about a car thief called Mikki. The article does not refer to Mikki as Romanian, although it does appear in the comments section. One commenter writes

He is not a Romanian! He is a GYPSY – from Romania. Gypsies came from INDIA few hundred years ago and we cannot get rid of them. They are doing the same job in Romania, they steal from **Romanians** also. We don't want them but we are forced to live with them. They cannot be integrated or sent to school. They are raising 6–7 a family and send them to beg. PLEASE CONTACT THE POLICE AND SEND HIM TO PRISON. THEY ARE A DISGRACE FOR THIS COUNTRY. 90–95 % OF THEM. Real **romanians** have **honest** jobs and they work very very hard to earn a

living, but this scumbags know only to steel, cheat, rob, bare knuckle fights. Dont judge Romania for this GYPSSYS. Ask a romanian about this GYPSSYS – they will tell you the same thing as me. And then ask brits about a native Romanian – you will hear good things.

Indeed, this comment contains other words which were identified as collocates of *Romanians* like *gypsys*. Furthermore, a closer look at Table 13.3 reveals collocates like *Roma* and *Indians* which also appear in comments which attempt to separate ‘honest, decent’ Romanians from the more negatively constructed gypsies; for example,

It is important to be informed before forming an opinion. Bulgarians and **Romanians** work, they don’t sleep on the streets. Only *gypsies* do as they don’t have dignity and don’t care where they live, as long as there is something to steal because this is what they call ‘work’.

(Comment 517)

This is a distinction which is made repeatedly in the comments section but does not appear to be so clear in the newspaper articles. The term *Roma* does occur in the articles, along with *gypsies*, although these terms are not frequent enough to be key. An illustrative case is article 45, which has the headline ‘Romanian gang ‘use fireworks to spark TERROR on Paris Metro – then pickpocket the victims’’. The body of this article does not use the word *Romanian(s)* again but instead refers to the gang as ‘a group of Roma women’ and ‘Roma pickpocket gangs’. From a critical perspective a number of issues arise here – the difference between Romanians and Roma gypsies, which appears to be highlighted in comments as more important than some newspaper articles appear to be acknowledging, and whether the more negative constructions of Roma gypsies are actually fair. However, at this point we will turn to look at some other keywords.

Pronoun Use

One difference between the two corpora was that the comments corpus contained several personal pronouns as keywords: *they*, *we*, *our*, *you*, *them* and *us*. To an extent this is to be expected as the commenting register is more colloquial and interactive than the news reporting register. However, pronouns can also have ideological functions; for example, they can help to create a shared sense of identity between a writer and reader, or they can construct in and out groups. For this reason, we have decided to examine two key pronouns, *them* (9873 occurrences in the comments corpus) and *us* (5944 occurrences in the comments corpus). We are particularly interested in the sorts of actions that are described as being carried out on these pronouns, so using the same settings for collocation as described earlier, Table 13.4 lists the verb collocates of each.

The verb collocates of *us* suggest a number of discourse prosodies (Stubbs 2001). The concept of discourse prosody relates to the ways that words can repeatedly co-occur in contexts which suggest a negative or positive evaluation. For example, Sinclair (1991: 112) has observed that ‘the verb *happen* is associated with unpleasant things – accidents and the like’. While the word *happen* does not appear to imply anything negative in itself, its prosody would indicate that people who encounter the word in text are likely to be primed to interpret what follows as likely to be negative.

Table 13.4 Verb collocates of *us* and *them* in the comments corpus

Chapter 6	Chapter 6
us	foisted, enslave, prevents, defraud, offload, forbids, outnumber, costing, betray, enlighten, warn, telling, inflicted, sponge, fleece, leech, deceive, betraying, dictate, treating, ripping, tells, denying, forgiven, enrich, store, bleed, hates, rob, dragging, bust, threaten, swamp, imposing, betrayed, sold, drag, tell, forcing, invade, imposed, forgive
them	carving, clothe, packing, shoving, bribe, send, rounding, offends, punish, escort, forgive, chuck, deport, turf, ship, educate, smuggle, shove, prosecute, educating, deporting, persuade, invite, shelter, shipping, sending, waving, encourages, handing, kick, lock

For the pronoun *us* in the comments corpus, the words *foisted*, *inflicted*, *imposing*, *forcing* and *imposed* suggest that some form of unwanted action has been carried out on ‘us’. These verbs refer to immigrants or EU-related laws which relate to immigration:

Don’t forget that we have ‘gained’ millions of unwanted and in many cases, worthless immigrants who were **foisted onto us** by the EU.

(123)

The dire effects of this forced multiculturalism **inflicted on us** by the cultural Marxists has already changed this country beyond recognition

(747)

Britons are seeing what has bee **imposed on us** from London and Brussels. ENOUGH IS ENOUGH,, NO MORE MIGRANTS.

Another (negative) discourse prosody involves burden, particularly relating to Romanians or immigrants in general, and involving verbs which operate as metaphors: *costing*, *sponge*, *fleece*, *leech*, *bleed*, *ripping*, *rob*:

80% of immigrants are just here to **fleece us** and moan in the process

(37)

we as a country are fighting for survival against this rabble who want to come here and **sponge of us**

(231)

Another mongrel immigrant to rob, rape, groom and **leech off of us**

(651)

They are coming here to **bleed us** dry because we are dull enough to give away FREE everything!

(203)

A third (negative) discourse prosody involves a sense of being overwhelmed or attacked: *outnumber, invade, swamp, enslave, threaten*:

Two faced Germans nothing changes they could not **enslave us** so they are buying us instead

(231)

The whole idea is to **swamp us** so that we lose our identity and they can rule us for ever.

(478)

And there is a sense of being coerced or lied to, not by Romanians but by those who are seen as elite decision makers: *deceive, betray, betraying, betrayed, defraud, dictate, forcing*:

And isn't it strange how our news channels never actually feature the depraved and disgusting behaviour of some of these immigrants? Presumably so our treacherous politicians can continue to lie, **deceive and betray us**, and keep pretending they are a great asset to our economy.

(617)

The EU serves to **dictate to us** what we can and can't do, whether we like it or not.

(14)

How about the verb collocates of *them*? The negative discourse prosody containing the most collocates involves some form of assisted movement: *shoving, send, rounding, escort, chuck, deport, turf, ship, smuggle, shove, deporting, invite, shipping, sending, waving and kick*. It is notable that these collocates do not appear to refer to Romanians specifically but more generally involve immigrants coming to the UK, particularly asylum seekers and undocumented migrants (this was also the case for the subject of the verb collocates earlier, like *sponge, bleed and fleece*).

Turf them out, **send them** back we are sick to death of these scroungers.

(193)

Ship them back to turkey regardless of where they claim to be from

(22)

The analysis of pronouns indicates that discussion of Romanians in the comments tends to take place within a wider context of immigration to the UK. This is borne out by reference to some of the other keywords in Table 13.1 (e.g. *migrants, immigration* and *immigrants*) are keywords in both corpora. The analysis so far reveals a fairly negative picture of how commenters view immigration generally and Romanians more specifically, as well as indicating how commenters create a shared sense of grievance, from a European political class who is seen to care more about immigrants than British nationals.

Identifying prototypical texts

This part of analysis takes a different approach by using a tool called ProtAnt which ranks a set of texts based on their lexical prototypicality. ProtAnt calculates the keywords in a

Table 13.5 The most prototypical articles in the corpus

	Headline	File number
1	Number of Romanians coming to Britain TRIPLES as EU workers boost migrant totals	152
2	We are losing control of our borders. Number of EU workers on the up	478
3	UKIP's Nigel Farage: 'Three-month benefit ban for migrants is not enough'	570
4	Number of Romanians and Bulgarians working in UK soars by nearly 10 PER CENT	381
5	Tory MPs warning over new tide of immigrants	1075
6	Romanians and Bulgarians were already working in Britain before ban ended	522
7	Up to 350,000 Romanians and Bulgarians coming to UK	970
8	Young Romanians want to work in the UK	907
9	New immigration BOMBSHELL as number of Romanians and Bulgarians in Britain TREBLE	515
10	Even Romanian MPs warn us that swarms of migrants are coming to live in UK	604

corpus of texts, using a reference corpus and then presents the list of texts in order, based on the number of keywords in each. Those at the top of the list are thus viewed as most typical. Experiments with ProtAnt, for example, Anthony and Baker (2015, 2016) showed that the tool was able to successfully identify the most typical newspaper articles and novels from corpora, as well as identifying atypical files (e.g. those which came from a different register to all the others in a corpus). Using ProtAnt with the corpus of 1929 *Express* articles of Romanians (and the BE06 corpus again used as the reference corpus) resulted in the articles in Table 13.5 being ranked as being most prototypical.

The headlines in Table 13.5 all point to articles about immigration to the UK, with seven of them directly referring to Romanians. From the perspective of an analyst carrying out CDA, it is also notable how metaphors are employed in some of the headlines. Article 1075 uses a water metaphor in referring to a *tide of immigrants*, whereas article 604 mentions *swarms of migrants*, figuratively constructing migrants as a kind of flying pest or insect.

One way that ProtAnt could be used is to carry out down-sampling of a larger dataset so that a close analysis could be carried out on a manageable number of typical files. For example, one of the files ranked in the ten most typical (file 515) is shown in full here.

New immigration BOMBSHELL as number of Romanians and Bulgarians in Britain TREBLE

BRITAIN was today hit with a new immigration bombshell as official figures showed the number of Romanians and Bulgarians arriving in the UK trebled in 2013.

Latest figures show net migration to Britain is rising again. The Office for National Statistics said this morning that 24,000 citizens from the two countries arrived in the year to September 2013, nearly three times the 9,000 who arrived in the previous 12 months. The ONS said this was 'statistically significant' and that around 70 per cent came to work, while 30 per cent came to study. Their figures also showed the Government was slipping behind its target for reducing overall net migration – the difference between migrants leaving and arriving in the UK. David Cameron wants the net number reduced to the tens of thousands but the figure soared to 212,000 in the period from 154,000 the previous year. Figures are not yet available for the numbers of

Romanians and Bulgarians who have arrived in the UK since January 1 when labour market restrictions were lifted to comply with EU rules. UKIP leader Nigel Farage said: 'The immigration targets were nothing more than spin to appease in the short term the public who are concerned about uncontrolled immigration. But how can you have any targets when you can't control who comes to live, settle and work in this country? It's as sensible as hefting butterflies.' Despite the sharp rise, Immigration and Security Minister James Brokenshire said a dip in the number of migrants arriving from outside the EU showed the Government was getting to grips with the issue. He said: 'Our reforms have cut non-EU migration to its lowest level since 1998 and there are now 82,000 fewer people arriving annually from outside the EU than when this government came to power. And overall figures are also well down from when we first came to government in 2010 – with nearly 70,000 fewer migrants coming to the UK. Numbers are down across the board in areas where we can control immigration, but arrivals from the EU have doubled in the last year.' Mr Brokenshire admitted the limitations of the Government's ability to restrict immigration from the EU. He said: 'We cannot impose formal immigration controls on EU migrants, so we are focusing on cutting out the abuse of free movement between EU member states and seeking to address the factors that drive European immigration to Britain.' A Home Office spokesman said reducing net migration to the tens of thousands 'remains the government's objective'.

A close qualitative analysis would focus on aspects of the language of this article (such as lexical choice, evaluation, metaphor and patterns around agency), as well as its text structure (for example, what pieces of information are foregrounded or appear at the end and are thus implied to be less important, who is quoted and how much space are they given and how does the main narrative voice of the article align itself with any quotes). For example, it is notable how in the headline the information about Romanians and Bulgarians is dramatically described as a *bombshell*. In the body of the article we are first given numerical information about numbers of immigrants, described as *official figures*, helping to legitimise the way in which the information is given. This is followed by a metaphorical description of how the government is *slipping behind* Prime Minister David Cameron's targets for immigration, which instead is said to have *soared*. Then the article notes that figures are not available since 1 January, an important date in this corpus as it was the date when Romanians and Bulgarians were allowed to travel to the UK to work. The lack of such information could implicitly invite readers to speculate on the amount of EU immigration since this date. After giving this information, there is a quote by the former leader of the UK Independence Party (UKIP), Nigel Farage, who describes the targets as 'spin'. The article then describes the rise in immigration as *sharp* and then gives a longer quote from a member of the government (James Brokenshire) who is directly responsible for immigration and attempts to downplay the amount of immigration to the UK. However, the minister is also described as admitting that the government are limited in being able to restrict immigration from the EU, and he refers to the existence of abuse of free movement between EU member states. The article then ends with a short statement from the Home Office about the objective of reducing net migration.

Although the language in the article is reasonably restrained in comparison to the language used in the comments, it is clear that immigration is taken for granted as problematic, with Britain described metaphorically as being *hit with an immigration bombshell*. The opinion of Nigel Farage is foregrounded with his quote appearing before those of government officials. Additionally, Farage is represented as speaking on behalf of the

whole country when he describes the public as being *concerned about uncontrolled immigration*. While the article appears to give a balanced view in that it presents opinions from members of the government which appear to be more reassuring about numbers of immigrants, the minister's speech appears to agree with Farage in terms of blaming membership of the EU for Britain's inability to restrict immigration. It is notable that nowhere in the article are any reasons given for why immigration to the UK is considered to be problematic. The public are described as 'concerned about uncontrolled immigration', but the article does not describe actual negative consequences of such immigration. This is left for readers to infer.

It is interesting to consider how the comments respond to the information in the article. The article produced 149 comments, of which over 95 percent express concern and disapproval about immigration to the UK. For example, commenters variously denounce immigrants as 'human shyte', 'scrounging load of rubbish', 'deluge of immigrants' and 'foreign spongers'. They are also implied to be criminal:

If we had a government in control of our country, not the puppets with their strings being pulled by the EU, we could deport all the foreign homeless, unemployed, beggars, pickpockets, rapists and murderers

The government is described as clueless, soppy, pretend, shambolic, powerless and useless. David Cameron is also viewed critically, for example, 'BALL – LESS apology for man', 'USELESS', 'A SICKENING SIGHT' and 'A WEAK SELF SERVING TRAITOR'. He is also described as 'clinging like a child to the skirt of Angela Merkel dictator of the EU'. On the other hand, UKIP is constructed as the only viable option, with 36 references to UKIP across the 149 comments, and 18 cases of the phrase *vote/voting UKIP*. A typical comment is 'The best for THIS Country is to have the UKIP in Power THEN we Will get a decient Britain like we used to have'.

Not all of the comments express anti-immigration and anti-EU views, with 5 percent offering an alternative perspective or are critical of the content of the article, for example,

Yesterday this paper reported that Romanians and Bulgarians returning to their countries because they were 'disillusioned with low pay and benefits' today the are apparently all coming back again- this has to be the most confused, dismal rag ever . . . Start reporting sensibly or I for one wont be buying this paper. You seem to be more confused than the govt is and that's saying something.

While such comments indicate that not everyone who reads the article is as accepting of the newspaper's stance, they are in the minority, and the alternative voices feel drowned out by the higher numbers of comments which not only agree with the article, but go further in terms of negatively constructing EU and British policy around immigration using much more offensive language. It is important to consider that the articles and comments therefore work in concert, presenting different but complementary perspectives. In a sense, the article initiates a discussion about Britain's role in Europe and the impact of immigration, putting forward a negative perspective but in a relatively 'restrained' tone. The comments continue this dialogue, using more inflammatory, emotive and colloquial language. They engage in 'decoding' work of the initial article, spelling out its implications, negatively characterising immigrants and the political establishment while advocating voting for UKIP, but in a minority of cases there is critical engagement giving an alternative perspective. It is by

considering both the articles and the comments in tandem that we can begin to understand the role that newspaper like *The Express* and its commentators played in the decision of the British people to leave the EU.

Future directions

A potential criticism of corpus linguistic approaches involves the application of cut-off points. For example, analysts need to determine a cut-off point for what counts as a keyword, and there are no clear guidelines on what this should be. While log-likelihood is a statistical test and thus associated p-values can be produced, the application of such tests on language data produces somewhat different results from the sort of hypothesis-testing experiments that are normally conducted in the social sciences. Thus, applying standard p-value cut-offs of 0.05 or 0.01 will often result in hundreds of keywords to examine. Also, the larger the corpus, the more keywords are likely to be produced. The notion of statistical significance is thus problematic and corpus analysts either impose very high p-values, like 0.0000001, or may simply select the strongest 10, 20 or 100 keywords, depending on how much analysis they are able to do. A good principle to bear in mind is transparency, so even if not all keywords are analysed, it is worth pointing out how many *could* have been analysed. For example, 1541 keywords were produced from the news articles which had a log-likelihood score of above 15.13 (corresponding to a p-value of <0.0001), while 1359 keywords were produced from the comments corpus using the same cut-off.¹⁹ Of the 50 which were examined from each corpus, they had a LL score of between 359 and 6470 in the articles and 728 and 7508 in the comments. Further research needs to be carried out on the application of keyword cut-offs but see Wilson (2012) for a discussion of how Zip's law might be gainfully used to this effect.

The consideration of user-generated data raises issues for corpus linguists. Readers will have probably noticed that the comments quoted in this chapter contain numerous typos. Such typos will impact on word frequency counts, and it is worth carrying out some preliminary research on the corpus to ascertain the extent to which this will impact on results. Tools such as VAARD (Archer et al. 2015) can be used to regularise inconsistent spelling in a dataset, although this, too, would require an initial survey of texts to outline the problem areas.

While CL methods are effective at handling large amounts of electronic text, they are currently not particularly good at dealing with other types of data (such as images, audio or video data). In order to analyse the relationship between image and text in a corpus of children's picture books, McGlashan (2016) developed a method called collustration which involved pairing concordance lines with images that occurred in the vicinity of the text under consideration. His analysis found how the images supplemented the text in interesting ways, for example, the phrase *love each other* tended to appear in the vicinity of images of people engaged in physical contact. However, this analysis needed to be carried out painstakingly by hand rather than automatically and would be difficult to implement on a large scale. Corpora could be encoded for images which are cross-linked to the parts of text that they appear next to and a future analysis tool could then show concordance lines alongside the corresponding images. Similarly, parts of the spoken British National Corpus have been cross-linked to audio files so that concordance lines can be clicked on to play audio clips of the speech. As new tools are developed, more sophisticated forms of analysis will become available.

In this chapter we have shown how corpus techniques can help to direct digital humanities analysts to salient or frequent linguistic patterns in large databases so that methods of

analysis associated with the kinds of close readings carried out by CDA can be utilised. Although we used a newspaper corpus (along with reader comments) as a case study, such techniques could be used on other electronic datasets of texts, including court proceedings, works of fiction, legal debates, diaries or business reports. The keywords procedure offered a ‘way in’ to the corpora we analysed, whereas the analysis of collocates enabled us to more quickly see repeated patterns of association with those words.

Finally, we note that CL techniques have a great deal to offer the digital humanities, although we advocate that they are used in conjunction with qualitative forms of inquiry which take various forms of context into account and involve close readings of texts. Otherwise, there is a danger in placing too much reliance on automatic forms of analysis, which may produce, at best, a shallow interpretation and, at worst, a completely inaccurate one.

Further reading

- 1 Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.
An essential text outlining practical approaches to corpus-based discourse analysis.
- 2 Baker, P., Gabrielatos, C. and McEnery, T. (2013). *Discourse analysis and media attitudes: The representation of Islam in the British Press*. Cambridge: Cambridge University Press.
This book brings together corpus linguistics and CDA to examine representations of Muslims and Islam in a large collection of UK news articles and provides a valuable example of how methods from CDA may be applied to very large corpora.
- 3 Baker, P. and McEnery, T. (eds.) (2015). *Corpora and discourse studies: Integrating discourse and corpora*. Basingstoke: Palgrave Macmillan.
A recent addition to the literature on and about the growing field of corpus-based discourse studies, this title presents a number of diverse approaches that might be taken when conducting corpus-based approaches studies of discourse in large datasets.
- 4 Wodak, R. and Meyer, M. (eds.) (2015). *Methods of critical discourse studies* (3rd ed.). Los Angeles: Sage Publications. Available at: <https://capitadiscovery.co.uk/bcu/items/1248946>.
A now-touchstone text for students of CDA. This accessibly written edited collection covers fundamental theory in CDA and presents several chapters from experts in the field on a range of theoretical and methodological approaches, including corpus linguistics.

Notes

- 1 www.theguardian.com/commentisfree/2016/jun/27/austerity-economic-woes-eu-referendum-brexite
- 2 www.dailymail.co.uk/news/article-3666511/Merkel-s-open-door-policy-caused-Britain-leave-EU-German-leader-blamed-Brexit-failure-deal-migrant-crisis-open-arms-immigration-policy.html
- 3 <http://lordashcroftpolls.com/2016/06/how-the-united-kingdom-voted-and-why/>
- 4 (www.dailymail.co.uk/news/article-2531440/Sold-Flights-buses-Romanians-Bulgarians-head-UK.html)
- 5 www.theguardian.com/media/greenslade/2014/mar/17/dailymail-pcc
- 6 We initially considered a number of possible newspapers, although found that each one had a different online commenting facility, which would have required a range of bespoke retrieval solutions, which this chapter would be unable to do justice to. The *Daily Express* worked well with the web scraping tools we wanted to showcase in this chapter.
- 7 www.python.org/
- 8 www.crummy.com/software/BeautifulSoup/ – BeautifulSoup is a library)
- 9 www.seleniumhq.org/
- 10 www.mozilla.org/en-GB/firefox/new/
- 11 www.google.com/chrome/

- 12 Further regarding ethics, it could be argued that when posting a comment on a freely available website which is not password protected, one's comments are already in the public domain. However, we felt it was good ethical practice to protect, as much as possible, the identities of commenters, especially as it would have been very difficult to ask permission of the hundreds of commenters to include their comments in the corpus. People who posted comments to the *Daily Express* website are required to create an anonymous account with a log-in name. With each comment only the log-in name is given with it – e.g. cheeky1952. We were unable to find references to real names in the dataset but to ensure an additional level of anonymity we do not refer to the log-in names in our analysis.
- 13 www.lexically.net/wordsmith/index.html
- 14 <http://corpora.lancs.ac.uk/lancsbox/>
- 15 <https://cqpweb.lancs.ac.uk>
- 16 www.sketchengine.co.uk
- 17 AntConc offers two ways of deriving keywords: chi-square tests and log-likelihood tests. There are currently at least eight ways of deriving keywords, some of which have strong champions and detractors; see <http://ucrel.lancs.ac.uk/llwizard.html>. It is beyond the remit of this chapter to do justice to the ongoing debate over which measure gives the 'best' set of keywords, but we would instead note that the log-likelihood test, which was used for this analysis, measures the confidence with which we can say that a given word is actually a keyword (e.g. the difference in frequencies is not the result of chance). However, the test does not give an indication of how strong the difference is.
- 18 The highest 50 LL scores range from between 330 and 3888 for the articles and 728 and 91407 for the comments. Having experimented with different cut-offs, (e.g. the top 10, the top 20, the top 50, the top 100), we felt that the top 50 keywords provided a reasonable number in order to allow keywords to be placed in categories of similarity to identify themes, without resulting in too much analytical repetition (e.g. resulting in lots of keywords which refer to the same theme). This is a matter of trial and error, and different cut-offs need to be established for each project, relevant with its aims and taking into account real-world practicalities.
- 19 <http://ucrel.lancs.ac.uk/llwizard.html>

References

- Anthony, L. (2016). AntConc (Version 3.4.4) [Computer Software]. Tokyo, Japan: Waseda University. Available at: www.laurenceanthony.net/
- Anthony, L. and Baker, P. (2015). *ProtAnt*: A tool for analysing the prototypicality of texts. *International Journal of Corpus Linguistics* 20(3): 273–292.
- Anthony, L. and Baker, P. (2016). *Evaluating the effectiveness of prototypical text detection in teaching and research: New developments and applications of ProtAnt*. Paper given at TaLC 2016 Gissen, Germany, 20–23 July.
- Archer, D., Kytö, M., Baron, A. and Rayson, P. (2015). Guidelines for normalising Early Modern English corpora: Decisions and justifications. *ICAME Journal* 39(1): 5–24.
- Baker, P. (2004). Querying keywords: Questions of difference, frequency and sense in keywords analysis. *Journal of English Linguistics* 32(4): 346–359.
- Baker, P. (2005). *Public discourses of gay men*. London: Routledge.
- Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics* 14(3): 312–337.
- Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McNery, T. and Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society* 19(3): 273–306.
- Baker, P., Gabrielatos, C. and McNery, T. (2013). *Discourse analysis and media attitudes: The representation of Islam in the British press*. Cambridge: Cambridge University Press.
- Chovanec, J. (2014). *The representation of Eastern Europeans in British internet debates*. Paper given at the 12th ESSE Conference, Košice, Slovakia, 29 August–2 September 2015.
- Durrant, P. and Doherty, A. (2010). Are high frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory* 6(2): 125–155.

- Fairclough, N. (1992). *Discourse and social change*. Cambridge: Polity Press.
- Fairclough, N. (1995). *Critical discourse analysis: The critical study of language*. London: Longman.
- Fairclough, N. (2003). *Analysing discourse: Textual analysis for social research*. London: Routledge.
- Fairclough, N., Wodak, R. and Mulderrig, J. (2011). Critical discourse analysis. In T. van Dijk (ed.), *Discourse studies: A multidisciplinary introduction* (2nd ed.). London: Sage, pp. 357–378.
- Fowler, R., Hodge, B., Kress, G. and Trew, T. (1979). *Language and control*. London: Routledge.
- Gabrielatos, C. and Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press 1996–2005. *Journal of English Linguistics* 36(1): 5–38.
- Gerbner, G., Gross, L., Morgan, M. and Signorielli, N. (1986). Living with television: The dynamics of the cultivation process. In J. Bryant and D. Zillman (eds.), *Perspectives on media effects*. Hilldale, NJ: Lawrence Erlbaum Associates, pp. 17–40.
- Habermas, J. (1988). *On the logic of the social sciences*. Cambridge, MA, and London: MIT Press.
- Jensen, K.E. (2014). Linguistics and the digital humanities: (computational) corpus linguistics. *Journal of Media and Communication Research* 30(57): 115–134.
- Kristeva, J. (1986). Word, dialogue and novel. In T. Moi (ed.), *The Kristeva reader*. Oxford: Basil Blackwell, pp. 34–61.
- McEnery, T., McGlashan, M. and Love, R. (2015). Press and social media reaction to ideologically inspired murder: The case of Lee Rigby. *Discourse and Society* 9(2): 1–23.
- McGlashan, M. (2016). *The representation of same-sex parents in children's picturebooks: A corpus-assisted multimodal critical discourse analysis*. PhD Thesis, Lancaster University, UK.
- Reisigl, M. and Wodak, R. (2001). *Discourse and discrimination: Rhetorics of racism and antisemitism*. London: Routledge.
- Sinclair, J. McH. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Stubbs, M. (2001). *Words and phrases*. London, Blackwell.
- Thurman, N. (2018). Newspaper consumption in the mobile age. *Journalism Studies* 19: 1–21.
- van Dijk, T. (1991). *Racism and the press*. London: Routledge.
- van Dijk, T. (2006). Discourse, context and cognition. *Discourse Studies* 8(1): 159–177.
- van Leeuwen, T. (1996). The representation of social actors. In C.R. Caldas-Coulthard and M. Coulthard (eds.), *Readings in critical discourse analysis*. London: Routledge.
- van Leeuwen, T. (1997). Representing social action. *Discourse and Society* 6(1): 81–106.
- Wenger, E. (1998). *Communities of practice: Learning, meaning and identity*. Cambridge: Cambridge University Press.
- Westlund, O. and Färdigh, M.A. (2015). Accessing the news in an age of mobile media: Tracing displacing and complementary effects of mobile news on newspapers and online news. *Mobile Media & Communication* 3(1): 53–74.
- Widdowson, H.G. (2004). *Text, context, pretext: Critical issues in discourse analysis*. Oxford: Blackwell.
- Wilson, A. (2012). Using corpora in depth psychology: A trigram-based analysis of a corpus of fetish fantasies. *Corpora* 7(1): 69–90.
- Wodak, R. and Meyer, M. (2009). Critical discourse analysis: History, agenda, theory and methodology. In R. Wodak and M. Meyer (eds.), *Methods of critical discourse analysis* (2nd ed.). London: Sage, pp. 1–33.

Conversation analysis

Eva Maria Martika and Jack Sidnell

Introduction

Conversation analysis is an approach to the study of social interaction which identifies and describes the stable practices of interaction and the encompassing organisations in which they are embedded. Its fundamental assumption is that naturally occurring talk is characterised by ‘order at all points’ (Sacks 1984), and this social order is to be found in the details of interactional events through detailed structural analysis of audio and video recordings of naturally occurring talk. Even though conversation analysis emerged within the field of sociology, it has predominantly focused on linguistic forms as a repertoire of practices for designing, organising, projecting and making sense of the trajectories and import of turns-at-talk. With its focus on linguistic objects as resources for constructing actions and sequences of actions in talk, the conversation analytic research enables us to understand some of the shaping factors of linguistic structures and patterns (grammar being the most researched).

The rapid development of digital technologies in the last two decades has implicated conversation analytic research in three possible ways: the digitisation of conversation analytic methods, the application of conversation analysis to text-based online interactions and the automation of conversation analysis. The goal of this chapter is to contribute to a growing body of literature in the digital humanities by addressing all the above three implications. While many definitions of the digital humanities ‘as smart and provocative as they are, often muddy the introductory waters more than clarify them’, and most definitions reduce the digital humanities to the ‘application of technologies to humanities work’ Gibbs (2013: 289), this chapter is in accordance with definitions that emphasize ‘studying the effects of the digital’ on human cultures as much as ‘using the digital’ to study human cultures (Gibbs 2013: 290-294).

The chapter is thus structured as follows: after a brief description of the foundational principles, basic methods and theoretical concepts of conversation analysis, we discuss the digitisation of conversation analytic research methods. We argue that while in adapting to the digital turn, conversation analysis retains its core methods for analysing interaction, digital technologies have enabled conversation analysts to better (and sometimes faster) answer long-standing questions and to pose innovative research questions that were difficult or even impossible to address without the use of digital means (e.g. searching for universal patterns in social interaction). Subsequently, we discuss conversation analytic studies of text-based online interactions, and we

report on some of the concerns and questions that conversation analysts have raised about the suitability of applying conversation analysis—a method developed for analysing face-to-face or telephone interaction—to text-based forms of talk on social media. We conclude this chapter by pointing out possible future directions in the field of conversation analysis such as, the expansion of the analytical work on embodied language and the automation of conversation analysis.

Background

Conversation analysis: foundations, premises and principles

The birth of conversation analysis is often attributed to a discovery that Harvey Sacks made while analysing phone calls to a suicide prevention line in Los Angeles in 1963 (Sacks 1967; Sidnell 2010: 11; Hutchby and Wooffitt 1998: 18; Psathas 1995: 13). One of the practical problems at the suicide prevention centre was that the call taker, generally a social worker, often found it difficult to elicit the name of the ‘suicidal’ caller. They, therefore, could not establish the identity of the caller, and this was considered crucial for the way in which they worked. While reviewing tapes of the phone calls to the suicide centre, Sacks attempted to answer the question of ‘Where in the course of the conversation could you tell that somebody would not give their name?’ (Sacks 1992: 3). It was this question and the subsequent examples of telephone openings to the suicide prevention line that particularly drew Sacks’s attention. For instance, Sacks considers the following two cases from his collections of examples:

1 Sacks, “Rules of Conversation Sequence”

[A= The social worker, B= The caller]
 A: This is Mr. Smith may I help you
 B: Yes, this is Mr. Brown

2 Sacks, “Rules of Conversation Sequence”

A: This is Mr. Smith may I help you
 B: I can’t hear you
 A: This is Mr. Smith.
 B: Smith

Examples such as these showed that one could ascertain possible difficulties in eliciting the caller’s name very early in the conversation. Sacks noted that if a staff member opened the phone call with ‘This is Mr. Smith may I help you’ the response would often be something like ‘Yes, this is Mr Brown’. If a different response was given and the name of the caller was not disclosed at the start of the phone call – as in example 2 – it would probably prove quite difficult to elicit the caller’s name in the rest of the call. Sacks also noted that you can obtain someone’s name without explicitly asking for it. Thus, one can simply say ‘This is Mr Smith’, and this will establish the relevance of the recipient giving his or her name, ‘This is Mr Brown’. This simple example shows that there are multiple ways of doing an action (e.g. asking someone’s name), and that there is a slot in conversations where a certain action is expected to occur. Sacks argued that if this slot is skipped, then, the action might never occur.

Just as there is a method of getting somebody’s name without asking for it, Sacks entertained the ‘wild’ possibility (Schegloff 1992: xvi-xvii) that the caller in example 2 could have methodically avoided giving their name (in response to the other’s having done so) by claiming a hearing problem. If this hypothesis was correct, this meant that talk could be

organised and designed at a high level of detail, thus the details of the turn-by-turn sequential organisation of interaction became the foci of attention in the emerging field.

From these and other initial observations (recorded in Sacks lectures), conversation analysis developed into an approach to the study of social interaction that emphasised detailed structural analysis and its fundamental assumption was that naturally occurring talk is characterised by ‘order at all points’ (Sacks 1984; Schegloff 1968; Sacks et al. 1974). Following in the path laid down by Sacks, Schegloff and Jefferson, the goal of conversation analysis has been to identify and to describe the stable practices of interaction and the encompassing organisations in which they are embedded. No interactional moment is ruled out as insignificant, random or irrelevant, because even the most mundane interactional activity can reveal an organisational sophistication if subjected to a systematic and detailed analysis (Atkinson and Heritage 1984: 4). Sacks (1984), Psathas (1990) and Schegloff (1996) suggest that an analyst should not have a narrow focus when analysing her data but should rather employ ‘unmotivated looking’ in the initial stages of research. In Psathas’s words, ‘the investigator is open to discovering [interactional] phenomena rather than instances of already identified and described phenomena or for some theoretically preformulated conceptualization of what phenomena should look like’ (1990: 24–25). This bottom-up analytical approach allows us to discover and understand more about the structure, the organisation and the subtle but important details of interaction (See Sidnell and Stivers 2013).

While the assumption about the orderliness of interaction has its roots in Goffman’s idea of interactional order (1983), the seeds from which conversation analytic methodology grew are to be found in Harold Garfinkel’s approach to the study of social phenomena. In accordance with ethnomethodological principles, Garfinkel developed conversation analysis out of an appreciation for ‘the common sense resources, practices and procedures’ through which interlocutors ‘produce and recognize mutually intelligible objects, events and courses of actions’ (Liddicoat 2007: 2).

The founders of conversation analysis sought to develop scientific, micro-analytic methods for understanding and describing the social order found in the details of interactional events. They developed a distinct line of work that analysed the procedures, practices, norms, and expectations through which interlocutors, *in situ*, achieve mutual intelligibility and which make orderly interaction possible. In the emerging methodology, the analytic focus was not on the content of interactional events or on ‘why people act as they do, but rather explicating *how* they do it’ (ten Have 2007: 9).¹

Sacks et al. (1974) paper on the turn-taking system provided an early demonstration of the method, in addition to explaining this domain of interactional organisation (See Schegloff 1968). The authors showed the methodical procedures the participants employ and jointly orient to in their efforts to achieve an orderly distribution and allocation of opportunities to produce a turn in a conversation.

While Sacks et al.’s paper on turn-taking was published in a linguistic journal (*Linguistic Society of America*), it adopted an approach to language quite different than the traditional linguistics at the time, which conceptualised linguistic categories and units as ‘static and pre-existing mental structures’ (Fox et al. 2013: 734). Instead, Sacks et al.’s analysis of the grammatical/syntactical typology and projectability of turn-construction units approached linguistic structure in terms of recurrent and stable conversational practices. Sacks et al. (1974: 723), conclude their paper remarking that

It is expectable . . . that some aspects of the syntax of a sentence will be best understood by reference to the jobs that need to be done in a turn-in-a-series, turns being a fundamental place for the occurrence of sentences.

According to Sacks et al., linguistic patterns are to be viewed, then, not as *a priori* categories, but as emerging *practices* tailored for specific interactional goals or tasks.

Critical issues and topics

Research in conversation analysis focuses on the details of how actions are implemented and recognised. The linguistic (and non-linguistic) objects used to design a turn are studied as resources to construct *actions* and *sequences of actions* in talk. The conversation analytic focus is on the challenges of what is often referred to as the problem of ‘action formation’ and ‘action ascription’ (Levinson 2013). Simply put, this is the question of how interlocutors use ‘the resources of the language, the body, the environment of the interaction, and position in the interaction’ in order to design and recognise particular social actions. Examples of social actions – out of a ‘class of unknown size’ (Schegloff 2007: xiv) – include requesting, inviting, complaining, agreeing, telling, rejecting and so on, as well as actions that do not have such familiar vernacular names but are nevertheless recognisable and describable.

Early conversation analysis emphasised that interlocutors rely on *turn position* and *turn composition* (grammatical format, lexical choices, prosody and other modalities) (Schegloff 2007) in order to recognise the particular action(s) implemented by each piece of interaction.² The composition or design of an utterance as a resource for understanding its meaning and import has received extensive attention in the field of linguistics, wherein the objects of analysis have been the lexical, morphological, syntactic, phonological and phonetic composition of phrases and utterances (Clift 2016: 32). However, conversation analysis considers ‘the *positioning* of an utterance in the ongoing conversation as fundamental to the understanding of its meaning and to the analysis of its significance as an action’ (Stivers 2013: 191). One of the central questions for the participants *and* for the conversation analyst trying to understand utterances in talk is ‘why that now’ (Schegloff and Sacks 1973: 299).

Focusing on action rather than meaning and taking into consideration the sequential position where something occurs is particularly helpful in analysing linguistic objects that do not have a straightforward meaning or a semantic core. Take, for example, the turn-initial particle ‘oh’ (Heritage (1984a, 1998). When used to preface the answer to a question, ‘oh’ can indicate that the answer is self-evident and the question need hardly have been asked (Heritage 1998). Consider the use of ‘oh’ in the following interview (analysed by Heritage 1998). Sir Harold Acton has been teaching poetry at Beijing University, and in lines 1–3, he mentions that his students were the first to translate T. S. Eliot’s work into Chinese (presumably, under his supervision). Given this information, the questioner (Harty) should have already assumed that Acton speaks Chinese. ‘Oh yes’ in line 6 marks the fact that the questions asks something that might have been presupposed from prior talk or background knowledge.

(A transcription key can be found at the end of the chapter.)

3 Chat Show

[Russell Harty- Sir Harold Acton]

01. Acton: . . hhhh and some of thuh- (0.3) some of my students
02. translated Eliot into Chine::se. I think thuh very
03. first.
04. (0.2)
05. Harty: Did you learn to speak (.) Chine[se.
06. Acton: -> [hh Oh yes.
07. (0.7)
08. Acton: .hhhh You ca::n’t live in thuh country without
09. speaking thuh lang[uage it’s impossible, hhhhh=
10. Harty: [Not no: course

Compare this with the use of ‘Oh’ in quite a different sequential position. When it prefaces the response to an answer to a question, ‘oh’ functions to register the answer to the question as informative. While in the first case ‘oh’ marks a shift from ‘what we are talking about’ to ‘background knowledge’, in this second case, ‘oh’ functions as a change-of-state token marking a change from not knowing to knowing (Heritage 1984a). The following example illustrates this. Notice in particular the way A explicitly articulates the change of state in her subsequent talk.

4 YYZ Washer-Dryer

01. A: . hh Do you guys have a washer en dry:er er something?
 02. (0.6)
 03.S: ah::=yeah, we got a little washer down here (.) °goin’°.
 04. A: -> o::h. ok. I didn’- I didn’t know that you guys
 ha:d that

We can see that the ‘meaning’ of oh is at least partially dependent on its sequential position. (See Heritage and Raymond 2005 for more uses of ‘Oh’.)

Sometimes, the non-occurrence of a linguistic item in a sequential environment can provide the analyst with evidence on its function in interaction. As Clift (2016: 17) remarks, ‘silence is as much an interactional object as the utterance preceding and following it’. To illustrate this, consider the following example from a doctor–patient interaction:

5 Frankel (1984)

- 01 Patient: This- chemotherapy (0.2) it won’t have any lasting
 02 effects on having kids will it?
 03 (2.2)
 04 It will?
 05 Doctor: I’m afraid so

In lines 1–2, the patient asks the doctor a polar question,³ which makes relevant a yes-no response, and which, by its design, conveys that its speaker believes the answer will be no (the chemotherapy will not have lasting effects). The absence of the doctor’s response, as indicated by the 2.2-second pause in line 3, is interpreted by the patient as problematic, and she subsequently produces a revised follow-up question in line 4: ‘It will?’ In contrast to the original question (line 1–2), the revised question (line 4) elicits an immediate response from the doctor (line 5). Conversation analysis is unique in attending to ‘those contexts of non-occurrences’ precisely because of its ‘focus on the online, incremental production of utterances in time’ (Clift 2016: 17).

In addition to the composition and the (sequential) position of a turn, other factors that play a role in action formation and ascription include the broader context of ongoing activities, the larger social framework, the social roles ascribed to participants and the epistemic statuses of the speakers (Heritage 2012). While other approaches in pragmatics and sociolinguistics take into consideration the social framework of the interaction and the contextual identities of the participants when analysing speech behaviour, conversation analysts do not assume that these aspects of social context will be procedurally consequential for the interaction.⁴ Rather, conversation analysis demands that we be able to show that some aspect of the context was in fact relevant to the participants in the course of their producing and interpreting the talk being analysed.

Consider in this respect the following example in which a father and a mother respond in very different ways to an observation made by a health visitor about their baby (line 1).

6 Heritage and Sefi (1992: 367)

(HV = Health visitor)

- 01 HV: He's enjoying that [isn't he.
 02 Father: [oYes, he certainly is=o
 03 Mother: =He's not hungry 'cuz(h)he's ju(h)st (h)had
 'iz bo:ttl ohhh
 04 (0.5)
 05 HV: You're feeding him on (.) Cow and Gate Premium.

Referring to some sucking or chewing that the baby is doing, the visitor remarks 'He's enjoying that isn't he'. The mother's and the father's responses (lines 2 and 3) set up two different contexts and ascribe two different actions to the health visitor's remark (line 1). While the father's agreement (line 2) treats the visitor's remark simply as an observation, the mother's response takes it as a challenge to her competence as a caregiver/mother. The mother responds with a defensive account of why the perceived criticism (that the baby may be hungry) cannot be valid. Thus, while the father's response constructs an everyday conversational context, the mother's response orients both 'to the institutional role of the health visitor as an observer and an evaluator of baby care and to her own responsibility and accountability for that care' (Drew and Heritage 1992: 33).

In this way, by grounding the analysis in the details of action formation, conversation analysis is able to understand how aspects of identity and the social roles of the participants are oriented to and negotiated as the interaction unfolds on a turn by turn basis rather than assuming 'some pre-established social framework . . . "containing" participants' speech behavior' (Clift 2016: 24)

While we have outlined some of the benefits of working with 'action ascriptions' in the conversation analytic tradition, there are some well-known problems with such an approach. First, categorising tokens as single, discrete acts is rarely a straightforward process because there is typically not a one-to-one mapping between practice and action (Sidnell 2010: 73). Second, ascribing actions to utterances often obscures the multiple, diverse things that participants do with the bit of conduct which is being placed 'under description' (Sidnell 2017: 322) in this way. Third, it is unlikely that participants in an interaction need to engage in such action ascription and categorisation in order to build a fitted response. While there are occasions in which participants do engage in explicit description (and thus imply categorisation), this involves rather a special form of interactional work that is by no means requisite to the orderly flow of conduct in an interaction (Sacks 1992). An alternative approach to the more familiar one based on action ascription involves an examination of the recurrent contextual factors that encourage the selection of one format versus others (Enfield and Sidnell 2017; Sidnell and Enfield 2017).

Main research methods

Data collection, transcription and analysis

Conversation analysis employs a bottom-up rather than a top-down approach to interactional phenomena. This approach implicates not only the analysis of the procedural aspects of interaction (as we saw in examples 1–5) but also the analysis of the social contingencies

within which the interactional episode occurs (as we saw in example 6). Due to space constraints, in this section, we briefly discuss only three of the conversation analytic research processes: data collection, transcription and the evidential basis of analysis.

The conversation analytic naturalistic stance to the study of interaction was evident in the methodological remarks of early publications in the field. For instance, Schegloff and Sacks (1973: 235) suggest that they undertake to ‘explore the possibility of achieving a naturalistic observational discipline that could deal with the details of social action(s) rigorously, empirically and formally’. Because of this naturalistic stance, conversation analysts rely on recordings of naturally occurring interaction rather than examples constructed by the researcher’s intuition, experiments, interviews or any other intervention by the researcher. Conversation analysts aim to capture the linguistic and embodied resources and the structural organisation of social activities in their ordinary environment, and they view interaction as being achieved incrementally through its temporal and sequential unfolding and collaboratively by all the participants involved. Also, as Sidnell (2010: 21) argues, conversation analysts insist on basing their analysis on recordings of actual talk because the level of complexity of such talk could neither be invented nor recollected (Sacks 1984).

In its approach to data collection, conversation analysis is different from linguistic fields such as semantics and most approaches in pragmatics, which usually employ imagined rather than naturally occurring examples. However, conversation analysis is similar to observational approaches such as sociolinguistics, interactional linguistics, linguistic anthropology or critical discourse analysis, which generally use recorded materials as their data. (For a detailed comparison between conversation analysis and these fields, see Clift 2016.)

Conversation analysis emerged at a time when linguistics was more occupied with idealised and abstract language structures than language *use*, and the details of everyday speech were considered too random, disorderly and degraded to be an object of inquiry (Heritage and Stivers 2013: 660; Clift 2016: 8–9). However, the outlook of conversation analysis was shaped in large part by Goffman’s conception of an interactional order, an idea born within a Durkheimian tradition, where interaction order is conceived as ‘existing prior to persons and their motivations, and indeed as structuring and sustaining them’ (Heritage and Stivers 2013: 663). Thus, the foundational premise of conversation analysis is that no interactional detail ‘can be dismissed *a priori* as disorderly, accidental, or irrelevant’ (Heritage 1984b: 242). This premise has implications for how conversation analysts go about working with data. First, the conversation analyst produces a transcription that retains as much detail as possible of the interactional event (Hutchby and Wooffitt 1998: 75). The goal is to transcribe not only *what* was uttered but also *how* it was uttered; the timing between the utterances, the interactants’ visible behaviours and background sounds are also transcribed. Second, the transcription needs to be structured in such a way as to accurately represent the interactional order the analyst aims to describe. The orderliness of the interaction needs to be readable as such in the transcript.

The transcription system most widely used by conversation analysts is the one produced by Gail Jefferson (Jefferson 1985, 2004; Hepburn and Bolden 2013: 57). (A transcription key can be found in the end of the chapter.) Jefferson’s system captures the sequential features of talk and is sensitive to the location of silence, onset of speech, overlapped speech, phenomena relevant to interactional units such as turns, turn transitions and turn completions, as well as more often overlooked elements (in other forms

of transcription) such as hitches, tokens and features of prosodic delivery (Psathas 1995: 70). Conversation analysts also use ‘adorned’ or ‘modified’ orthography whenever the speakers deviate significantly from pronunciations that the standard orthography is meant to represent (Walker 2013: 471).

While the goal of the transcription is to capture the conversational event as it actually occurs, ‘transcripts are necessarily selective in the details that are represented, and thus, are never treated by conversation analysts as a replacement for the data’ (Hepburn and Bolden 2013: 57). Repeated listening and viewing of the recordings during the production of a transcript enables the conversation analyst to locate the recurring patterns of different conversational phenomena.

The basic form of evidence in conversation analysis is the ‘next-turn proof procedure’ (Sidnell 2010: 67, 2013: 79; Sacks et al. 1974). This involves tracking the participants’ own displayed orientations and understandings as evidence to support the analysis of what a particular practice is doing. Consider one case to exemplify this form of evidence and compare it to other analytical approaches in linguistics. The following example is from a phone call between two female neighbours, Nancy and Emma. Emma has tried to reach Nancy earlier on the phone, but this has not been possible because Nancy’s phone has been busy for some time. The extract starts with Emma remarking that Nancy’s phone line has been busy.

7 Pomerantz (1980: 195)

- 01 Emma: .Yer line’s been busy.
 02 Nancy: Yeuh my fu(hh).hh my father’s wife called me. .hh
 So when she
 03 calls me::, .hh I always talk fer a long time.
 Cuz she c’n afford
 04 it’n I can’t .uhhh heh .ehhhhhh

If we were to only search for the meaning of Emma’s utterance (‘Your line has been busy’, line 1), we might argue that she is simply informing, observing or asserting something. However, Nancy’s response (lines 2–4) constitutes evidence that she is not orienting to this semantic interpretation of Emma’s utterance. By giving an extensive account of *why* her line was busy, Nancy’s response treats the declarative ‘Yer line’s been busy’ as seeking rather than simply sharing information. Based on a collection of such cases, Pomerantz (1980) described this practice as ‘fishing’ because it involves a speaker eliciting information from a recipient simply by telling of their limited access to some set of events which the recipient knows about. In other words, the telling of an experience can serve as a possible elicitor of information. So, in the case earlier, Emma’s report of having noticed that Nancy’s line as busy provides for Nancy (the recipient) to voluntarily disclose some information (i.e. the reason why the line has been busy, who she was talking to, etc.) (See Enfield and Sidnell 2017.)

This analysis is not tied to the grammatical form of Emma’s utterance – a declarative. Pomerantz (1980) did not argue that a declarative by itself acts as a fishing device, nor that a declarative always implements ‘fishing’. Rather, she argued that this practice is tied to the epistemic asymmetry⁵ that exists between Emma and Nancy and to the sequential position in which the telling occurs (e.g. as a first pair part).

Digitisation of conversation analytic research methods

Given the significant development of digital technologies since the early 2000s, how do digital technologies impact conversation analytic research methods, premises and findings? We will argue that digital technologies have opened up new possibilities for what conversation analysis can achieve by capturing, storing, analysing and representing digital data.

Following we discuss how the digitisation of some conversation analytic research methods has facilitated one of the foundational analytical principles of conversation analysis – that is, the ‘unmotivated looking’.⁶ As Ehrlich and Romaniuk (2014: 471) explain:

The term is meant to imply that the analyst be open to discovering what is going on in the data rather than searching for a particular pre-identified or pre-theorized phenomenon. For conversation analysts, careful and repeated listening to (and viewing of) recorded interaction in transcribing data and producing a transcript constitutes an important initial step in the process of data analysis. Indeed . . . transcription operates as an important ‘noticing device’.

Sophisticated tools for video, audio, image and text analysis allow for less time consuming and, in some cases, more precise ‘noticing devices’, as well as for collections of readily playable instances of the phenomenon under analysis. Mainstream media player applications and multimedia frameworks such as Windows Media Player, RealPlayer, QuickTime and VLC allow conversation analysts to view and listen to their recordings as well as precisely select and crop sections of interest. Particularly helpful in the transcribing process, though, have been professional software applications such as list-format transcribers – Computerized Language Analysis (CLAN)⁷ and Transana⁸ – and multimedia annotators and ‘partition format’⁹ transcribers such as ELAN¹⁰ and PRAAT.¹¹ The two formats enable different representations of the data. The list-format transcriber is better for the representation of sequence organisation, while the partition format works better for the representation and analysis of a multimodal interaction, where various lines of action need to be simultaneously represented and analysed (Albert 2012).

The digitised transcribing process has not only increased the *quantity* of the ‘noticings’ but also improved their *quality* by enabling validation (of the transcribed material) and the comparison of the (recorded and transcribed) ‘noticed’ instances with the click of a button.

A conversation analytic domain that has relied significantly on digital technology is research on *embodied action*. A key finding in conversation analysis research is that ‘various features of the delivery of talk and other bodily conduct are basic to how interlocutors build specific actions and respond to the actions of others’ (Hepburn and Bolden 2013: 55). Conversation analytic research on embodied action has relied significantly on digital technology not only in order to produce a detailed transcription and analysis of the temporal and sequential relationship between audible and visible aspects of social action but also to represent these aspects in a more ‘holistic’ and ‘interpretable’ way.¹² As Hepburn and Bolden (2013: 70) have pointed out, ‘digital frame grabs can be selected to capture, with a high degree of precision, the temporal unfolding of visible conduct in coordination with speech’ and ‘visual representations (e.g. drawings and digital video frames) accompanying a transcript of vocal conduct have the advantages of being easily interpretable and more holistic in representation’.

Along the same lines, Heath and Luff (2013: 306) argue that the technological developments such as the shift from analogue to digital video

have not only provided us with the resources to be able to record and see aspects of activities that hitherto were largely inaccessible, but to edit and combine recordings, for example, from multiple cameras, that hitherto would have been severely challenging if not impossible.

Heath and Luff see the digital impact on video-based conversation analytic research not only ‘in terms of the range of phenomena, activities, even settings, that can be subject to analytic scrutiny, but also with regard to how insights and observations can be revealed and presented’.

Current contributions and research

Conversation analytic research with digital technologies

The digitalisation of conversation analysis has aided in the task of distinguishing between universal and cultural-specific aspects of human interaction. Conversation analytic methodology was initially applied to recordings of English data.¹³ Later on, conversation analytic studies shed light on the interactional particularities and linguistic resources in languages other than English. For the past 25 years, there has been a growing body of conversation analytic work on a broad range of languages such as Japanese (Fox et al. 1996; Hayashi 1999, 2003), German (Egbert 1996; Golato and Fagyal 2008), Yele¹⁴ (Levinson 2005), Mandarin (Wu 2004) and French (Chevalier 2008). (See Sidnell 2007 and Dingemanse and Floyd 2014 for an overview.) These studies have demonstrated how ‘generic interactional issues or “problems” (e.g. how turns are to be distributed among participants) are solved through the mobilization of local resources (grammar, social categories, etc.)’ in various languages and communities (Sidnell 2007: 229). Broad-ranging and large-scale¹⁵ quantitative conversation analytic studies across multiple languages have only become a possibility in the last decade, with the development of new digital tools (from the software mentioned earlier to other information-sharing technologies such as emails, Dropbox, etc).

The dissemination of data and insights in a digital way, as well as the possibility for distributional analysis, has facilitated the search for universals and cultural variation in conversational structures and practices (Stivers et al. 2009). For example, Stivers et al.’s (2009) paper on universals and cultural variation in turn-taking compared question–response sequences in video recordings of informal natural conversation across ten languages, from five continents. Conversation analytic methods were used to identify and characterise the phenomena to be coded, while digital technologies were crucial in the comparative analysis of the coded phenomena. Stivers et al. (2009: 10591) coded response times – that is, the time elapsed between the end of the question turn and the beginning of the response turn – both in the traditional conversation analytic way (that is, auditorily)¹⁶ and through the annotation software ELAN.¹⁷ Computational methods were employed to calculate the ‘mean, median, and mode information for each language’, to examine the distribution of turn offset times across the ten languages, to test the factors that account for faster or slower response times in these different languages and to generate comparative measures (Stivers et al. 2009: 10591).¹⁸

In addition to cross-linguistic comparative studies, in the last two decades, large corpus-based conversation analytic studies have documented correlational and distributional

relationships between interactional practices and social outcomes by digitally processing large collections of institutional data mainly in the field of medicine (Heritage and Stivers 1999, Haakana 2002; Stivers and Majid 2007; Heritage et al. 2007). For example, Heritage et al. (2007) studied the effects of a doctor's question design on whether a patient felt their medical concerns were sufficiently met during a medical visit. They found that in cases of patients with multiple concerns, disclosing secondary concerns was significantly encouraged or discouraged by the doctor's word selection in asking the patients to share these. Specifically, a question such as 'Is there *anything* else you want to address in the visit today?' was found to be twice as likely to receive a negative response versus questions asked using 'something' (that is, 'Is there *something* else you want to address in the visit today?'). Heritage et al. (2007) argued that choosing to ask the patient with the negative polarity term 'anything' of the first question tilted the preference towards a negative response.¹⁹ In other words, by conveying an expectation that the patient does not have *anything* else to share, doctors discouraged the patients from articulating their concerns.

At this point we need to clarify that while conversation analytic findings can provide an empirical foundation for quantitative studies on the frequency or distribution of interactional practices and their correlation with various (social) outcomes, statistical relationships are not within the methodological premises, goals and assumptions of conversation analysis. In fact, conversation analysis emerged within Goffman's vision of furthering our 'power of observation' (Maynard 2013: 15), which draws attention to in-depth analysis of face-to-face interaction and was introduced as an alternative to the methods of survey and statistical analysis (Maynard 2013: 15). Arguing for the interactional functions or consequences of a practice or a format in conversation analysis does not depend on statistical validity, but on the participants' own displayed orientations to an interactional norm. Sometimes, even a single instance can inform conversation analysts about the existence of a conversational norm. As Fox et al. (2013: 735–736) have pointed out 'the "regularity" of the format is not achieved probabilistically or statistically, but rather normatively, and participants orient to deviation from the format'. For conversation analysts, generating claims that are empirically defensible means that such claims can be shown to be both relevant to the participants and procedurally consequential for the interaction. Thus, the statistical inquiries mentioned earlier rely on conversation analytic research methods to identify practices to be coded. Conversation analysis provides the 'operational definitions of the interactional practices that will be coded and counted', while statistical research 'addresses different types of research questions, involves different data-gathering techniques, and employs different modes of analysis' (Gill and Roberts 2013: 590).

Conversation analysis of online interaction: methodological developments and challenges

Even though conversation analysis developed as a method for analysing co-present or phone interactions, it has been employed as a method for analysing text-based online interactions in at least 89 peer-reviewed articles since the 1990s (Paulus et al. 2016: 1). The application of conversation analysis to online data begs several questions: for instance, how similar or different are co-present and text-based forms of talk? How suitable is conversation analysis for analysing online data? What challenges does this application of conversation analysis entail? What foundational structures of conversation, as they have been described by conversation analysts, are also found in the analysis of online data? Two recent articles address

these questions and provide an extensive and helpful overview of conversation analytic-informed research on online talk: Paulus et al.'s (2016) literature review of studies employing conversation analysis as a method for analysing online data and Giles et al.'s (2015) introduction to the methodological development of 'digital' conversation analysis. Following we summarise some of their main points.

According to the literature review conducted by Paulus et al. (2016), conversation analytic studies on online interaction have focused on four main areas: (1) comparing online and face-to-face or phone conversations, (2) understanding how coherence is maintained in online interaction, (3) understanding how participants deal with interactional troubles and (4) understanding how social actions are accomplished asynchronously (Paulus et al. 2016: 1).

Online talk takes place through a range of modalities, and conversation analysis has been employed to analyse talk produced in 'quasi'-synchronous chat/messaging, asynchronous blog/forum posts and comments, as well as emails and social media posts/comments (Paulus et al. 2016: 4). While the distinction between synchronous and asynchronous talk becomes more and more problematic as the modalities for online communication evolve in various forms, researchers have retained this distinction when describing their online data (Paulus et al. 2016: 4). This is probably because, as Paulus et al. (2016: 5) assert in their literature review, while asynchronous data can be 'vastly different . . . from spoken talk', most researchers have concluded that (quasi-synchronous) 'text-based online chat looked much like voice-based F2F or telephone conversations, with some effect due to the conversational medium' (Paulus et al. 2016: 5).

Similar to conversation analytic studies of offline talk, the studies analysing online talk also orient 'to more than one conversational structure as part of their analysis' (Paulus et al. 2016: 4). According to Paulus et al. (2016), the most frequent structures of conversation taken up by conversation analysis-based studies of online talk are sequence organisation, turn design, turn allocation/turn-taking/turn-construction units (TCUs), membership category analysis,²⁰ repair (see next paragraph), preference structure and storytelling.

Meredith and Stokoe's (2014: 181) study on *repair* practices in quasi-synchronous Facebook 'chats' demonstrates the analytical complexities of a conversation analytic approach to offline talk. Repair is 'an omnirelevant feature of interaction', and any aspect of talk can become a target for repair, from word selection to prosody (Meredith and Stokoe 2014: 184). Examples of troubles are 'misarticulations, malapropisms, use of a "wrong" word, unavailability of a word when needed, failure to hear or to be heard, trouble on the part of the recipient in understanding, incorrect understandings by recipients' (Schegloff 1987: 210). Repair can be initiated by the speaker or the recipient of the trouble source. The first would be a case of self-initiated repair, while the latter would be a case of other-initiated repair.

Meredith and Stokoe's (2014: 201) analysis of self-repair instances in online interaction showed that self-initiated self-repair 'occurs in online interaction, in both similar and contrasting ways to repair operations in spoken conversation'. So, for example, they found that what they call the 'visible' form of repair occurs in similar sequential positions as in spoken interaction, and the preference for self-initiated self-repair is evident in online interaction too (2014: 202). On the other hand, they identified a form of repair that is not available in spoken interaction: the 'message construction' repair (2014: 202). This form of repair occurs before the message is sent to the recipient, and is thus unavailable to the recipient. This means that, unlike in spoken interaction, the online recipient does not have access to action formation efforts such as restarting or reformulating one's turn, and the troublesome

bits of online conduct are not accountable as they would be in spoken interaction (2014: 202). However, even this form of online repair displays similarities with the spoken interaction. As Meredith and Stokoe (2014: 181) argue,

while the practice of message construction repair is made possible through the affordances of the online medium, it nevertheless shows how participants in written interaction are oriented to the same basic contingencies as they are in spoken talk: building sequentially organized courses of action and maintaining intersubjectivity.

Overall, Meredith and Stokoe (2014: 181) consider any conclusion about the differences between spoken and online interaction as premature, and they propose that ‘online talk should be treated as an adaptation of an oral speech-exchange system’.

According to Paulus et al.’s literature review, recent publications find the narrow comparison between online and offline talk inadequate, and they propose transcending the methodological tools, concepts and structures used to analyse offline talk (2016: 5). Lamerichs and te Molder (as cited in Paulus et al. 2016: 5) have proposed analysing online interactions as ‘social practices in their own right’. Gonzalez-Lloret (2011: 313), who analysed strategies used to maintain coherence in online talk, has suggested that ‘rather than imposing existing structures on the new medium, we should examine the ways in which participants achieve different sequence types in the new medium’. Unsurprisingly, these criticisms and proposals are consistent with the basic principle of ‘unmotivated looking’ in conversation analysis, which, as mentioned earlier, requires the analyst to be open to new structures and phenomena rather than seeking to confirm pre-existing ones.

Along the same lines, Giles et al. (2015: 45–46) established the Microanalysis of Online Data (MOOD) network in order to develop ‘conversation analytic-informed methodologies and methods to carry out research on new media with the same degree of intellectual rigor that has been applied to the study of offline talk-in-interaction’. Despite this scope, their ‘interest is with the dynamics of online communication *per se*, and [their] view is that online communities are interesting in and of themselves, not as weak simulacra of offline communities’ (Giles et al. 2015: 47). Advocating the development of a ‘digital’ conversation analysis, Giles et al. (2015: 50) have called for ‘ongoing methodological discussions’ in order to ‘refine the application of conversation analysis to online talk’.

Giles et al. (2015: 46) acknowledge that figuring out whether existing methods can be applied uncritically to online data or whether it is necessary to create new methods tailored to the different types of online data is a complicated task. However, they support the view that

the move from an uncritical ‘digitized’ application of conversation analysis to a customized version of conversation analysis for specific use with online interaction requires the reworking of a number of tenets of conversation analysis in the light of the challenges posed by electronic communication technology

(Giles et al. 2015: 47)

Data transcription is one of the methodological practices that is affected.

The transcription process which, as mentioned earlier, is, in traditional conversation analysis, a ‘noticing device’, is almost eliminated altogether when analysing online data. The researcher from the beginning works with a transcript produced by the interlocutors themselves. While this reduces the degree of mediation between the transcript and the interlocutors, it requires a re-appropriation of the traditional conversation analysis transcription

system (Jefferson 2004). For example, overlaps in online interactions do not occur in a similar way with offline interactions. Thus, their representation needs to be modified. Also, most of the conventions for representing prosodic delivery are made redundant in online interactions (e.g. the symbols for indicating changes of pitch and volume – except, perhaps, for the capital letters representing loud talk). Conventions for in-breaths and laughter particles are also made redundant. Having explored how transcription methods might function in the contexts of online data, Meredith and Potter (2014) have called for new forms of transcribing and have supported the view that data in conversation analytic studies of online talk should include screen recordings of synchronous chat participants' real-time interactions.

Future directions

For over 45 years, conversation analytic research has focused on the generic organisation of talk, and it has extensively described the orderly practices of social action in interaction. While such research has shed light into the fine-grained organisation of interactional structures such as sequence and turn-taking, it has also enabled us to understand some of the shaping factors of linguistic structures and patterns (grammar being the most researched). Conversation analysis has focused on grammatical forms as a repertoire of practices for designing, organising, projecting and making sense of the trajectories and import of turns-at-talk. Thus, as Thompson and Couper-Kuhlen (2014: 161) argue, 'recurrent regularities of turn organization and turn sequences . . . are shown by interactional linguists to be the primary motivations for the development of grammatical patterns'.

Conversation analytic research, consistent with this assumption that linguistic patterns are emergent phenomena arising from everyday interactions, has more specifically provided evidence that 'linguistic regularities are dependent on, and shaped by, not just the context in which they occur, but more specifically by the actions which the speakers are undertaking with their talk' (Thompson and Couper-Kuhlen 2014: 161). As displayed earlier, talk is organised around sequences of actions, which are collaboratively achieved on a turn-by-turn basis, and conversation analytic research has shown that conversationalists systematically rely on linguistic patterns to build these sequences of actions. They do so when they project the completion of a turn, the type of action(s) under way, etc. The choice of one grammatical form (among more than one alternative) for accomplishing an interactional task informs us about the function as well as the import of a grammatical form (Curl and Drew 2008).

Even though conversation analysis emerged within the field of sociology, it has significantly interacted with linguistics. Very often conversation analytic courses are taught in a language and linguistics departments, and articles on conversation analysis have been published in a number of linguistics journals. The contributions of conversation analysis to linguistics are nicely summarised by Fox et al. (2013: 729):

Conversation analysis has brought to linguistics fresh perspectives on data, methods, concepts and theoretical understandings. . . . Conversation analysis has given linguistics an enriched view of data and transcription, bringing attention not just to everyday conversation but to its minute details of its moment-by-moment unfolding. Conversation analysis has deepened the appreciation for the home of many linguistic practices in everyday talk and has offered a temporally dynamic approach to language . . . shift in thinking about the linguistic form as static to a view of linguistic patterns as practices fitted to particular sequential environments.

The methodological and theoretical benefits of conversation analysis together with the development of digital technologies have expanded the employment of conversation analysis as a research method in various fields such as education, medicine, legal process, human–machine interaction and software engineering (Heritage 2010b). However, morpho-syntactical phenomena have predominantly been the focus of conversation analytic investigation, while work on embodied interactional practices and prosody is still in its infancy. As we discussed earlier, digital technologies can support conversation analysis research in expanding this line of work.

The analytic developments in conversation analysis and digital technologies have even generated ambitions for a potential automation of conversation analytic methods. Automation of conversation analysis is an emerging research area, and some recent work has explored the automation of diagnostic conversation analysis in patients with memory complaints (Mirheidari et al. 2017) and the potential of automated transcription technology for navigating big conversational datasets (Moore 2015), to mention just a few. The automation of conversation analysis though faces certain challenges as it relies on the successful methods of a few more disciplines: automated speech recognition, speaker diarisation (who is speaking when), spoken language understanding, etc. (Mirheidari et al. 2017: 2). Crucially, automation of conversation analysis runs into the same old problem of ‘action ascription’ mentioned in the first section. As discussed earlier, interlocutors and conversation analysts rely on multiple clues – linguistic and non-linguistic items, sequential position, turn design, the social environment of the interaction, etc. – in order to recognise what is accomplished in every bit of interaction. Automation of conversation analysis raises further doubts about its possibility to pay attention to the subtleties of conversation such as the ones revealed by the analysis of examples 6 and 7 – that is the orientation of participants to different social identities and epistemic (a)symmetries.

Scholars in digital humanities (Burdick et al. 2012) have argued that ‘digital capabilities have challenged the humanist to make explicit many of the premises on which those understandings are based in order to make them operative in computational environments’. While conversation analysis has made pretty explicit its premises, Fröhlich and Luff (1990) warn us against the dangers of misunderstanding the foundations of conversation analysis and misapplying its findings. Button (1990) discusses three such cases of misunderstanding in the application of conversation analytic findings to human–computer interaction. Other scholars find conversation analysis a more promising approach to artificial intelligence and human–computer interaction than cognitive science (Luff 1990), and they argue that conversation analysis can help in efforts to emulate human interaction and in the design of human–computer interactions. There is already a growing body of research in studying different conversational structures. (For the study of repair in human–computer interaction see Raudaskoski 1990 and in human–computer interaction or interactions with smart assistants in home settings such as Amazon Alexa see Velkovska and Moustafa 2018, or Opferman 2019.).

APPENDIX: TRANSCRIPTION CONVENTIONS

- [A left bracket indicates the onset of overlapping speech
-] A right bracket indicates the point at which overlapping utterances end
- = An equals sign indicates contiguous speech
- (0.2) Silences are indicated as pauses in tenths of a second
- (.) A period in parentheses indicates a hearable pause (less than two tenths of a second)
- . A period indicates a falling intonation contour
- , A comma indicates continuing intonation
- ? A question mark indicates rising intonation contour
- _ An underscore indicates a level intonation contour
- : Colons indicate lengthening of preceding sound (the more colons, the longer the lengthening)
- A hyphen indicates an abrupt cutoff sound (phonetically, a glottal stop)
- but Underlining indicates stress or emphasis, by increased amplitude or pitch
- BUT Upper case letters indicates noticeably louder speech
- °oh° The degree sign indicates noticeably quiet or soft speech
- hh The letter 'h' indicates audible aspirations (the more hs the longer the breath)
- .hh A period preceding the letter 'h' indicates audible inhalations (the more hs the longer the breath)
- y(h)es h within parentheses within a word indicates "laugh-like" sound
- ~ A tilde indicates tremulous or 'wobbly' voice
- (guess) Words within single parentheses indicate likely hearing of that word
- ((coughs)) Information in double parentheses indicate descriptions of events rather than representations of them
- () Empty parentheses indicate hearable yet indecipherable talk

Notes

- 1 The ethnomethodological understandings underlying conversation analysis have often led some to refer to it as 'ethnomethodological conversation analysis' (Schegloff et al. 2002: 3). This type of conversation analysis is what we discuss in this chapter, and the early work of this conversation

- analysis is associated with scholars such as Sacks, Schegloff, Jefferson, Pomerantz, Psathas, Heritage and others. This conversation analysis is not to be confused with the ‘conversational’ analysis literature produced by scholars such as Grice, Gumperz, Tannen and some scholars from the field of applied linguistics (Schegloff et al. 2002: 3).
- 2 Conversation analysts, too, rely on these same resources in order to analyse and understand what (actions) participants are doing with each bit of interaction.
 - 3 Specifically, this is a negative declarative with a positive tag.
 - 4 For a detailed comparison between the conversation analytic and pragmatic or sociolinguistic approaches to the social context of interaction, check Clift 2016.
 - 5 As Heritage (2012) explains, there is an epistemic asymmetry here between the two because Nancy knows why her line has been busy, whereas Emma does not.
 - 6 It is important to clarify here that digital technology has simply facilitated the basic principle of ‘unmotivated looking’; it has not enabled it. Conversation analysts were doing ‘unmotivated looking’ with analogue data as well.
 - 7 CLAN is a free software developed by the CHILDES and Talkbank project.
 - 8 www.transana.com/
 - 9 This descriptive term was found in Saul Albert’s blog on transcribing tools for conversation analysis research: <http://saulalbert.net/blog/conversation-analytic-transcription-with-clan/>
 - 10 ELAN is a free tool developed by the TLA at Max Planck Institute for the creation of complex annotations on video and audio resources. (www.lat-mpi.eu/tools/elan)
 - 11 Praat is a free software used to make a phonetic analysis of speech. For more information, you can check www.fon.hum.uva.nl/praat/
 - 12 See, for example, Levinson (2005: 438–439) where he presents figures of sequences of still taken from video data, as they are processed on ELAN.
 - 13 Moerman’s (1977) work on Thai being the only exception.
 - 14 A ‘Papuan’ language spoken in Rossel Island, an island off the eastern tip of Papua New Guinea.
 - 15 ‘Large-scale’ comparative research in the conversation analysis tradition varies from 10 (Stivers et al. 2009) to as many as 20 languages (Enfield et al. 2013). The rigorous conversation analysis processes of transcribing, analysing individual instances and building collection of instances puts limitations on the number of languages that can be studied.
 - 16 This is a subjective and relative way of measuring the elapsed time. The conversation analyst in this case measures the time between the turns by taking into consideration the overall pace of the interaction.
 - 17 In this case, the elapsed time was measured in 10-ms increments.
 - 18 This multivariate analysis was performed by using STATA, a statistical software package.
 - 19 See Pomerantz and Heritage 2013 for an overview of the conversation analytic concept ‘preference’.
 - 20 Schegloff’s (2007) tutorial on membership categorisation outlines the major components of membership categorization devices (MCDs) and their implications for research practice in conversation analysis and social sciences in general.

Further reading

- 1 Meredith, J. and Stokoe, E.H. (2014). Repair: Comparing facebook ‘chat’ with spoken interaction. *Discourse and Communication* 8(2): 181–207.
This paper is an excellent example of how conversation analytic methods and concepts can be applied and appropriated to analyse online conversations, and it reveals some primary findings from a conversation analysis–based comparison between online and offline talk.
- 2 Paulus, T., Warren, A. and Lester, J.N. (2016). Applying conversation analysis methods to online talk: A literature review. *Discourse, Context and Media* 12, June: 1–10.
This is a concise review of 89 studies applying conversation analysis on online talk, and it outlines some of their main areas of focus, as well as their limitations or challenges.
- 3 Sacks, H., Schegloff, E.A. and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation *Language* 50(4): 696–735.
This is a foundational paper in conversation analysis constituting the very first effort to exemplify the orderliness of conversation by describing its most basic structure: turn-taking.

- 4 Sidnell, J. (2010). *Conversation analysis: An introduction*. Oxford: Wiley-Blackwell.
This book explains the crucial domains of organisation in conversation, and it introduces the main findings, methods and analytical techniques of conversation analysis. It also provides an engaging historical overview of the field.
- 5 Sidnell, J. and Stivers, T. (eds.) (2013). *The handbook of conversation analysis*. Chichester, West Sussex, UK: Wiley-Blackwell.
This book provides a comprehensive review of the key methodological and theoretical underpinnings of conversation analysis. It engages in a discussion with an extensive body of international research in the field for the past 45 years, it pinpoints the latest developments and critical issues in the field, it describes the relationship of conversation analysis with other disciplines and it suggests future directions for research in the field.

References

- Albert, S. (2012). Conversation analytic transcription with clan. *Saul Albert*, 4 October. Available at: <http://saulalbert.net/blog/conversation-analytic-transcription-with-clan/> (Accessed 26 June 2019).
- Atkinson, J.M. and Heritage, J. (eds.) (1984). *Structures of social action: Studies in conversation analysis*. Cambridge: Cambridge University Press.
- Burdick, A., Ducker, J., Lunenfeld, P., Presner T. and Schnapp, J. (2012). *Digital humanities*. Cambridge: MIT Press.
- Button, G. (1990). Going up a blind alley. In P. Luff, N. Gilbert and D. Fröhlich (eds.), *Computers and conversation*. London: Academic Press, pp. 67–90.
- Chevalier, F. (2008). Unfinished turns in French conversation: Projectability, syntax and action. *Research on Language and Social Interaction* 41(1): 1731–1752.
- Clift, R. (2016). *Conversation analysis* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.
- Curl, T.S. and Drew, P. (2008). Contingency and action: A comparison of two forms of requesting. *Research on Language and Social Interaction* 41(2): 129–153.
- Dingemanse, M. and Floyd, S. (2014). Conversation across cultures. In N.J. Enfield, P. Kockelman and J. Sidnell (eds.), *Cambridge handbook of linguistic anthropology*. Cambridge: Cambridge University Press, pp. 434–464.
- Drew, P. and Heritage, J. (1992). Analysing talk at work: An introduction. In P. Drew and J.C. Heritage (eds.), *Talk at work*. Cambridge: Cambridge University Press, pp. 3–64.
- Egbert, M. (1996). Context-sensitivity in conversation: Eye gaze and the German Repair Initiator *bitte?* *Language in Society* 25: 587–612.
- Ehrlich, S. and Romaniuk, T. (2014). Discourse analysis. In R.J. Podesva and D. Sharma (eds.), *Research methods in linguistics*. Cambridge: Cambridge University Press, pp. 460–493.
- Enfield, N.J., Dingemanse, M., Baranova, J., Blythe, Brown, J.P., Dirksmeyer, T., Drew, P., Floyd, S., Gipper, S., Gisladdottir, R.S., Hoymann, G., Kendrick, K., Levinson, S.C., Magyari, L., Manrique, E., Rossi, G., San Roque, L. and Torreira, F. (2013). Huh? What? – A first survey in 21 languages. In G. Raymond, J. Sidnell and M. Hayashi (eds.), *Conversational repair and human understanding*. Cambridge: Cambridge University Press, pp. 343–380.
- Enfield, N. and Sidnell, J. (2017). *The concept of action*. Cambridge: Cambridge University Press.
- Fox, B., Hayashi, M. and Jaspersen, R. (1996). Resources and repair: A cross-linguistic study of syntax and repair. In E. Ochs, E.A. Schegloff and S.A. Thompson (eds.), *Interaction and grammar*. Cambridge: Cambridge University Press, pp. 185–237.
- Fox, B.A., Thompson, S.A., Ford, C.E. and Couper-Kuhlen, E. (2013). Conversation analysis and linguistics. In J. Sidnell and T. Stivers (eds.), *The handbook of conversation analysis*. Chichester, West Sussex: Wiley-Blackwell, pp. 726–740.
- Frankel, R.M. (1984). From sentence to sequence: Understanding the medical encounter through micro-interactional analysis. *Discourse Processes* 7: 135–170.

- Fröhlich, D.M. and Luff, P. (1990). Applying the technology of conversation to the technology for conversation. In P. Luff, G.N. Gilbert and D.M. Fröhlich (eds.), *Computers and conversation*. London: Academic Press, pp. 237–260.
- Gibbs, F. (2013). Digital humanities. Definitions by type. In M. Terras, J. Nyhan and E. Vabhoutte (eds.), *Defining digital humanities: A reader*. Farnham: Ashgate.
- Giles, D., Stommel, W., Paulus, T., Lester, J. and Reed, D. (2015). Microanalysis of online data: The methodological development of “digital Conversation Analysis”. *Discourse, Context and Media* 7: 45–51.
- Gill, V.T. and Roberts, F. (2013). Conversation analysis in medicine. In J. Sidnell and T. Stivers (eds.), *The handbook of conversation analysis*. Chichester, West Sussex: Wiley-Blackwell, pp. 575–592.
- Goffman, E. (1983). The interaction order: American sociological association, 1982 presidential address. *American Sociological Review* 48(1): 1–17.
- Golato, A. and Fagyal, Z. (2008). Comparing single and double sayings of the German response of the German response token *ja* and the role of prosody: A conversation analytic perspective. *Research on Language & Social Interaction* 41(3): 241–270.
- Gonzalez-Lloret, M. (2011). Conversation analysis of computer-mediated communication. *CALICO Journal* 28(2): 308–325.
- Haakana, M. (2002). Laughter in medical interaction: From quantification to analysis, and back. *Journal of Sociolinguistics* 6(2): 207–235.
- Hayashi, M. (1999). Where grammar and interaction meet: A study of co-participant completion in Japanese conversation. *Human Studies* 22: 475–499.
- Hayashi, M. (2003). *Joint utterance construction in Japanese conversation*. Amsterdam: John Benjamins.
- Heath, C. and Luff, P. (2013). Embodied action and organizational activity. In J. Sidnell and T. Stivers (eds.), *The handbook of conversation analysis*. Chichester, West Sussex: Wiley-Blackwell, pp. 283–306.
- Hepburn, A. and Bolden, G.B. (2013). The conversation analytic approach to transcription. In J. Sidnell and T. Stivers (eds.), *The handbook of conversation analysis*. Chichester, West Sussex: Wiley-Blackwell, pp. 57–76.
- Heritage, J. (1984a). A change-of-state token and aspects of its sequential placement. In J.M. Atkinson and J. Heritage (eds.), *Structures of social action*. Cambridge: Cambridge University Press, pp. 299–345.
- Heritage, J. (1984b). *Garfinkel and ethnomethodology*. Cambridge: Polity Press.
- Heritage, J. (1998). Oh-prefaced responses to inquiry. *Language in Society* 27(3): 291–334.
- Heritage, J. (2010a). Questioning in medicine. In A. Freed and S. Ehrlich (eds.), *Why do you ask. The function of questions in institutional discourse*. New York: Oxford University Press, pp. 42–68.
- Heritage, J. (2010b). Conversation analysis: Practices and methods. In D. Silverman (ed.), *Qualitative sociology* (3rd ed.). London: Sage, pp. 208–230.
- Heritage, J. (2012). The epistemic engine: Sequence organization an territories of knowledge. *Research on Language and Social Interaction* 45(1): 30–52.
- Heritage, J. and Raymond, G. (2005). The terms of agreement: Indexing epistemic authority and subordination in talk-in-interaction. *Social Psychology Quarterly* 68(1): 15–38.
- Heritage, J., Robinson, J.D., Elliott, M.N., Beckett, M. and Wilkes, M. (2007). Reducing patients’ unmet concerns in primary care: The difference one word can make. *Journal of General Internal Medicine* 22(10): 1429–1433.
- Heritage, J. and Sefi, S. (1992). Dilemmas of advice: Aspects of the delivery and reception of advice in interactions between health visitors and first time mothers. In P. Drew and J.C. Heritage (eds.), *Talk at work*. Cambridge: Cambridge University Press, pp. 359–419.
- Heritage, J. and Stivers, T. (1999). Online commentary in acute medical visits: A method of shaping patient expectations. *Social Science & Medicine* 49(11): 1501–1517.
- Heritage, J. and Stivers, T. (2013). Conversation analysis in sociology. In J. Sidnell and T. Stivers (eds.), *Handbook of conversation analysis*. Boston: Wiley-Blackwell, pp. 659–673.

- Hutchby, I. and Wooffitt, R. (1998). *Conversation analysis: Principles, practices and applications*. Cambridge: Polity Press.
- Jefferson, G. (1985). An exercise in the transcription and analysis of laughter. In T.A. Dijk (ed.), *Handbook of discourse analysis*, vol. 3. New York: Academic Press, pp. 25–34.
- Jefferson, G. (2004). A note on laughter in ‘male-female’ interaction. *Discourse Studies* 6(1): 117–133.
- Levinson, S. (2013). Action formation and ascription. In J. Sidnell and T. Stivers (eds.), *The handbook of conversation analysis*. Chichester, West Sussex: Wiley-Blackwell, pp. 103–130.
- Levinson, S.C. (2005). Living with Manny’s dangerous idea. *Discourse Studies* 7: 431–453.
- Liddicoat, A. (2007). *An introduction to conversation analysis*. London: Continuum.
- Luff, P. (1990). Introduction. In P. Luff, G.N. Gilbert and D.M. Fröhlich (eds.), *Computers and conversation*. London: Academic Press, pp. 237–260.
- Maynard, W.D. (2013). Everyone and no one to turn to: Intellectual roots and contexts for conversation analysis. In J. Sidnell and T. Stivers (eds.), *The handbook of conversation analysis*. Chichester, West Sussex: Wiley-Blackwell, pp. 11–31.
- Meredith, J. and Potter, J. (2014). Conversation analysis and electronic interactions: Methodological, analytic and technical considerations. In H.L. Limand and F. Sudweeks (eds.), *Innovative methods and technologies for electronic discourse analysis*. Hershey, PA: IGI Global, pp. 370–393.
- Meredith, J. and Stokoe, E.H. (2014). Repair: Comparing facebook ‘chat’ with spoken interaction. *Discourse and Communication* 8(2): 181–207.
- Mirheidari, B., Blackburn, D., Harkness, K., Walker, T., Venneri, A., Reuber, M. and Christensen, H. (2017). Toward the automation of diagnostic conversation analysis in patients with memory complaints. *Journal of Alzheimer’s Disease* 58(2): 373–387.
- Moerman, M. (1977). The preference for self-correction in a Thai conversational corpus. *Language* 53(4): 872–882.
- Opferman, C.S. (2019, July). *Users’ adaptation work to get Apple Siri to work: An example of structural constraints on users of doing repair when talking to a ‘conversational’ agent*. Paper presented at the IIEMCA19, Mannheim, Germany.
- Paulus, T., Warren, A. and Lester, J.N. (2016, June). Applying conversation analysis methods to online talk: A literature review. *Discourse, Context and Media* 12: 1–10.
- Pomerantz, A. and Heritage, J. (2013). Preference. In J. Sidnell and T. Stivers (eds.), *The handbook of conversation analysis*. Chichester, West Sussex: Wiley-Blackwell, pp. 210–228.
- Pomerantz, A.M. (1980). Telling my side: ‘Limited access’ as a ‘fishing device’. *Sociological Inquiry* 50: 186–198.
- Psathas, G. (1990). Introduction: Methodological issues and recent developments in the study of naturally occurring interaction. In G. Psathas (ed.), *Interactional competence*. Washington: University Press of America, pp. 1–30.
- Psathas, G. (1995). *Conversation analysis. The study of talk-in-interaction*. London: Sage.
- Raudaskoski, P. (1990). Repair work in human – computer interaction. A conversation analytic perspective. In P. Luff, N. Gilbert and D. Fröhlich (eds.), *Computers and conversation*. London: Academic Press, pp. 151–171.
- Robert, J.M. (2015). Automated transcription and conversation analysis. *Research on Language and Social Interaction* 48(3): 253–270.
- Sacks, H. (1967). The search for help: No one to turn to. In E. Schneidman (ed.), *Essays in self-destruction*. New York: Science House, pp. 203–223.
- Sacks, H. (1984). On doing ‘being ordinary’. In J.M. Atkinson and J. Heritage (eds.), *Structures of social action: Studies in conversation analysis*. Cambridge: Cambridge University Press, pp. 413–429.
- Sacks, H. (1992 [1964]). Lecture 1: Rules of conversational sequence. In G. Jefferson (ed.), *Harvey sacks, lectures on conversation. (Fall 1964 – Spring 1968)* (Vol. 1). Oxford: Blackwell, pp. 3–11.
- Sacks, H., Schegloff, E.A. and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50(4): 696–735.
- Schegloff, E.A. (1968). Sequencing in conversational openings. *American Anthropologist* 70(6): 1075–1095.

- Schegloff, E.A. (1987). Analyzing single episodes of interaction: An exercise in conversation analysis. *Social Psychology Quarterly* 50(2): 101–114.
- Schegloff, E.A. (1992). Introduction. In G. Jefferson (ed.), *Harvey sacks, lectures on conversation (Fall 1964 – Spring 1968)* (Vol. 1). Oxford: Blackwell, pp. ix – lxii.
- Schegloff, E.A. (1996). Confirming allusions: Toward and empirical account of action. *American Journal of Sociology* 102(1): 161–216.
- Schegloff, E.A. (2007). *Sequence organization*. Cambridge: Cambridge University Press.
- Schegloff, E.A. and Sacks, H. (1973). Opening up closings. *Semiotica* 8: 289–327.
- Schegloff, E.A., Koshik, I., Jacoby, S. and Olsher, D. (2002) Conversation analysis and applied linguistics. *Annual Review of Applied Linguistics* 22: 3–31.
- Sidnell, J. (2007). Comparative studies in conversation analysis. *Annual Review of Anthropology* 36: 229–244.
- Sidnell, J. (2010). *Conversation analysis: An introduction*. Oxford: Wiley-Blackwell.
- Sidnell, J. (2013). Basic conversation analytic methods. In J. Sidnell and T. Stivers (eds.), *The handbook of conversation analysis*. Chichester, West Sussex: Wiley-Blackwell, pp. 191–209.
- Sidnell, J. (2017). Action in interaction is conduct under a description. *Language in Society* 46(3): 313–337.
- Sidnell, J. and Enfield, N. (2017). Deixis and the interactional foundations of reference. In Y. Huang (eds.), *Oxford handbook of pragmatics*. Oxford: Oxford University Press, pp. 217–239.
- Sidnell, J. and Stivers, T. (eds.) (2013). *The handbook of conversation analysis*. Chichester, West Sussex: Wiley-Blackwell.
- Stivers, T. (2013). Sequence organisation. In J. Sidnell and T. Stivers (eds.), *The handbook of conversation analysis*. Chichester, West Sussex: Wiley-Blackwell, pp. 191–209.
- Stivers, T., Enfield, N.J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoyman, G., Rossano, F., de Ruiter, J.P., Yoon, K. and Levinson, S.C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences* 106(26): 10587–10592.
- Stivers, T. and Majid, A. (2007). Questioning children: Interactional evidence of implicit bias in medical interviews. *Social Psychology Quarterly* 70: 424–441.
- ten Have, P. (2007). *Doing conversation analysis: A practical guide* (2nd ed.). London: Sage.
- Thompson, S. and Couper-Kuhlen, E. (2014). Language function. In N. Enfield, P. Kockelman and J. Sidnell (eds.), *The Cambridge handbook of linguistic anthropology*. Cambridge: Cambridge University Press, pp. 158–182.
- Velkovska, J. and Moustafa, Z. (2018). The illusion of natural conversation: Interacting with smart assistants in home settings. In *CHI'18*. Montreal, Canada.
- Walker, G. (2013). Phonetics and prosody in conversation. In J. Sidnell and T. Stivers (eds.), *The handbook of conversation analysis*. Chichester, West Sussex, UK: Wiley-Blackwell, pp. 455–473.
- Wu. (2004). *Stance in talk: A conversation analysis of Mandarin final particles*. Amsterdam: John Benjamins.

Cross-cultural communication

Eric Friginal and Cassie Dorothy Leymarie

Introduction

Digital technology has greatly increased researchers' ability to collect and analyse large amounts of linguistic data. Language as the primary element of human communication holds many answers to questions of how to better facilitate effective interaction across cultures in both spoken and written forms (Connor 1996; Tannen 2000). The digital humanities have enabled researchers to create and use electronic tools to enhance quantitative and qualitative approaches to understanding communication and how it functions across multiple settings (Belcher and Nelson 2013). An application of these approaches is amplified in more specialised sub-fields such as computational and corpus linguistics, digital recording and photography *vis-à-vis* language ecology and, especially, linguistic landscape and computer-assisted language analysis.

Kirschenbaum (2014) refers to the digital humanities as a social undertaking that brings together 'networks of people' (p. 56). Likewise, Terras (2016) acknowledges that digital humanities studies make use of 'computational methods that are trying to understand what it means to be human, in both our past and present society' (p. 1638). Thus, when cultures interface, the analysis of language, and in particular its textual and visual forms, become of utmost importance. Clearly, there are many other interrelated approaches to highlight here, but this chapter describes two prominent methods in analysing cross-cultural communication and understanding in the digital humanities: (1) corpora and corpus linguistic approaches and (2) digital photography and linguistic landscape studies. While corpus linguistics is a more established approach in quantifying language for facilitating better communication, the study of the cross-cultural linguistic landscape is still relatively in its infancy. However, linguistic landscape is a method that shows much promise for counting and indexing language use in multilingual societies, creating 'digital cultural artefacts' that can be analysed through computing and interpreting active and immersive cultural symbols that may represent both the present and the past (Terras 2016).

The focus of this chapter, therefore, is to overview current research in cross-cultural communication that utilises these technologies and digital humanities tools and approaches

in a variety of settings. Specifically, the contributions of corpora and corpus linguistics and digital photography are overviewed and highlighted. Studies from various disciplines concerned with language and communication are outlined across contexts, including (1) the use of corpora and corpus tools in understanding cross-cultural communication in the workplace (e.g. healthcare, office communication and telephone-based business interactions) and the use of digital photography in understanding everyday language from signage in multilingual communities. A sample analysis focusing on corpora in cross-cultural telephone-based business interactions is featured here. Key issues and the important contributions of digital technologies are presented and discussed, focusing on how research and data from the aforementioned settings can enhance our contextual understanding of cross-cultural communication and potentially support the acquisition of cross-cultural competence.

Corpora and cross-cultural communication

Background

Corpus linguistics has become increasingly popular as a research approach that facilitates practical investigations of language variation and use, producing a range of reliable and generalisable linguistic data on the characteristic features of cross-cultural communication that can be extensively interpreted. The corpus linguistic approach makes use of methodological innovations that allow scholars to ask research questions across many cross-cultural settings. The findings these questions generate produce important perspectives on language variation from those taken in traditional, ethnographic studies (Biber et al. 2010). Corpora have provided a strong support for the view that language use in cross-cultural communication is systematic, yet nuanced, and can be described more extensively using empirical, quantitative methods. One important contribution of the digital humanities to communication studies, therefore, is the documentation of the existence of cultural constructs that may strongly (i.e. statistically) influence language variation and use.

Corpora in exploring cross-cultural communication

The word *corpus* has been used in digital humanities to refer to datasets comprising written and (transcribed) spoken texts (McEnery et al. 2006). Instead of just being used to describe any collection of information, corpora are seen as systematically collected, naturally occurring samples of texts. With personal computers, the Internet, and the digitisation of everyday language, various corpora – both general and specialised – have become more widely available and shared (Sinclair 2005). Anyone with a computer and access to the Internet can begin to create a corpus, as long as it follows a logical and linguistically principled design. It is important to emphasise that a logical corpus design attempts to fully represent the range of target language features in spoken and written texts necessary for users to arrive at sound conclusions and interpretations. Over the years, the number of publicly available online corpora continues to increase exponentially, which researchers around the world can study and analyse (Friginal and Hardy 2014).

Corpus linguistics is not, in itself, a model of language but is primarily a methodological approach that can be summarised according to the following considerations (Biber et al. 1998: 4):

- It is empirical, analysing the actual patterns of use in natural texts
- It utilises a large and principled collection of natural texts, known as a corpus (pl. corpora), as the basis for analysis
- It makes extensive use of computers for analysis, employing both automatic and interactive techniques
- It relies on the combination of quantitative and qualitative analytical techniques

It is important to remember that although corpora offer measurable, frequency-based descriptions of texts and cultural groups (depending on the nature and composition of the corpus), the researcher and subsequent readers of these studies must still interpret these corpus-informed findings qualitatively using nuanced interpretive techniques. Hence, corpus methodologies are often utilised *in tandem* with register-based methods in interpreting data based on contextual information, role relationships, target audience and other situational characteristics. This means that frequency and statistical distributions from corpora have to be functionally and accurately interpreted. In an interview with Friginal (2013a), Biber emphasised the importance of functional interpretation of corpus-based data:

Quantitative patterns discovered through corpus analysis should always be subsequently interpreted in functional terms. In some corpus studies, quantitative findings are presented as the end of the story. I find this unsatisfactory. For me, quantitative patterns of linguistic variation exist because they reflect underlying functional differences, and a necessary final step in any corpus analysis should be the functional.

(p. 119)

Corpora and corpus tools provide various options to answer a wide range of questions in digital humanities and several other branches of linguistics and communication studies. Corpus-based analysis is ‘a methodology that uses corpus evidence as a repository of examples to expound, test or exemplify given theoretical statements’ (Tognini-Bonelli 2001: 10). The development and now relatively easy access to computational tools (e.g. concordancers, part-of-speech taggers and parsers, linguistic data visualisers) that readily process huge volumes of texts make it possible to investigate prominent discourse characteristics and compare their distributions across various corpora and/or communicative domains.

Current contributions and research (and primary research methods)

Corpora and corpus linguistics research approaches

Corpus-based studies have explored the patterns and functions of extensive distributions of the lexical and grammatical features of written and spoken registers of many languages, especially English. Corpus findings have gradually explained and incrementally defined the functional parameters of language based on frequencies and statistical patterns of usage (Pickering et al. 2016). Consequently, some multilingual and cross-cultural corpora have been developed and collected, and many scholars have utilised corpus-based data and approaches in their own research comparing linguistic behaviours and norms of cultural groups.

National corpora such as the British National Corpus (BNC), the American National Corpus (ANC), and online language varieties or registers collected and shared by Davies from Brigham Young University (www.corpus.byu.edu) all contain several millions of electronically stored spoken and written texts that are routinely used as resources for a range

of applications, from language teaching, to dictionary production, to sociolinguistic studies, in order to research natural language processing. These online resources also feature various speakers and writers representing a range of cultural groups. Text extraction tools and concordancing software such as Wordsmith Tools (Scott 2010) and AntConc (Anthony 2014), which enable users to obtain distributions of words and phrases, are easily accessible or downloadable online (note that AntConc is freeware). Many other computational programmes that allow complex tagging and annotations of text and audio and video data are now also quite familiar to a growing number of scholars and researchers. These include the ELAN platform from the Max Planck Institute (https://tla.mpi.nl/tools/tla-tools/elan/citing_elan/), the currently very popular and consistently updated Sketch Engine programme (www.sketchengine.co.uk/) and the Stanford Parser (<http://nlp.stanford.edu/software/lex-parser.shtml>), which has been one of the primary sources for natural language processing (NLP) codes and algorithms. There is also continuing interest in the development of more specialised corpora designed to address specific contexts such as academic spoken and written English (the Michigan Corpus of Academic Spoken English [MICASE] and the Michigan Corpus of Upper-Level Student Papers [MICUSP]) or English as a lingua franca (the Vienna – Oxford International Corpus of English [VOICE]) (Friginal et al. 2017; Pickering et al. 2016).

In sum, developments in corpus collection, matched by the expansion of corpus-based approaches, have positioned researchers in digital humanities to study cross-cultural communication using a range of numerical data. The common quantitative designs that focus on the lexico-grammatical features of different registers from frequency counts now also include the examination of keyword lists; p-frames, n-grams and lexical bundles; syntactic complexity indices; and multidimensional analyses (e.g. Belcher and Nelson 2013; Biber 1995; Friginal 2009b; Staples 2016). Corpus-based discourse studies also apply methods of qualitative discourse analysis to examine pragmatic and cross-cultural questions which may then be *supplemented* by conversation analysis, linguistic profiling, critical discourse analysis and more in-depth qualitative coding of themes and communicative markers. Mixed-method approaches could then be utilised with frequency-based data to further interpret various discourse features of speakers/writers in cross-cultural settings as they perform various communicative tasks.

Corpora and studying cross-cultural workplaces

The number of studies focusing on frequency-based linguistic and pragmatic features of the intercultural workplace has increased significantly over the past several years. In addition, qualitative analyses of spoken interactions, based on recorded data of professional discourse (e.g. by exploring moves, turn-taking or repair and action formulation), have been used over the years to explore various implications of a given utterance relative to the grammar of spoken discourse and the influence of speakers' cultural background and awareness during the interaction (Belcher and Nelson 2013). With corpora in the digital humanities, several qualitative and quantitative studies have been published, merging mixed methodologies and providing more generalisable conclusions. Specialised corpora in the professional fields have been collected, primarily with the focus on improving our understanding of professional communication and to develop training materials to enhance service quality, avoid miscommunication and facilitate successful business transactions. Some of the most commonly studied registers of cross-cultural workplace discourse using corpora include office-based workplace interactions (e.g. Di Ferrante 2013); outsourced call centre transactions

(e.g. Friginal 2009b, 2013b); healthcare, specifically nurse–patient interactions (e.g. Staples 2016); international business (e.g. Warren 2014); and tourism, hotel and service industry communications (e.g. Cheng and Warren 2006; Warren 2014), among many others.

Of note here is the Hong Kong Corpus of Spoken English (prosodic) (HKCSE) which features a variety of formal and informal office talk, customer service encounters in hotels, business presentations and conference telephone calls. As a cross-cultural corpus, the primary cultural groups communicating in many of the workplaces are Chinese speakers from Hong Kong and native and non-native English speakers from many different countries. The HKCSE was collected between 1997 and 2002 and includes a sub-corpus of business English of approximately 250,000 words (Cheng et al. 2008; Warren 2004). This corpus collection is also manually annotated for prosodic features using a model of discourse intonation developed by Brazil (1997). A built-in concordancing programme called iConc was developed for the HKCSE and allows quantitative analyses of annotated intonational features (Cheng et al. 2006). In addition to discourse intonation, many interrelated studies, as well as subsequent specialised corpora collected by the research team in Hong Kong, have specifically explored the lexical and grammatical features of transactional dialogues, keywords and keyness values of business English; and pragmatic markers of transactional negotiations, realisations of intertextuality, directionality and hybridisation in the discourse of professionals; and politeness strategies in disagreements (Cheng and Warren 2006; Warren 2011, 2013, 2014; Pickering et al. 2016).

The Language in the Workplace Project (LWP) corpus is a national corpus collected by scholars from Victoria University in Wellington, New Zealand. The LWP includes a wide range of largely office-based workplaces in government organisations and small business settings in New Zealand. Other locations include hospitals, publishing companies, information technology companies and news organisations. An extensive amount of research has been undertaken using the LWP since the late 1990s ranging from book-length manuscripts to conference papers (Pickering et al. 2016). Specifically, the LWP corpus has been used to explore cross-cultural pragmatics, speakers' gender and ethnicity and language use in the workplace, humour, small talk and speech acts, such as directives and requests (Holmes 2006; Marra 2012; Stubbe et al. 2003). Two recent studies by Vine (2016, 2017) using the LWP explore the use of the pragmatic markers *you know*, *eh* and *I think*; and *actually*, *just* and *probably* in office-based interactions using a theory of cultural dimensions (Hofstede 2001) to locate New Zealand workplaces on a continuum of power and formality (from informal conversations to formal unscripted monologues). Vine reported that these pragmatic markers occur in both larger, more formal meetings as well as short, informal dialogues. Although the use of *probably* did not vary between the two contexts, *actually* and *just* occurred more frequently in the small, less formal interactions.

For business-specific cross-cultural research in British and American English, corpora such as the Cambridge and Nottingham Business English Corpus (CANBEC) and the American and British Office Talk Corpus (ABOT) have been explored across a variety of linguistic comparisons. CANBEC is a 1-million-word sub-corpus of the Cambridge English Corpus (CEC) covering business settings from large companies to small firms and both transactional (e.g. formal meetings and presentations) and interactional (e.g. lunchtime or coffee room conversations) language events. Studies using CANBEC have focused on the distribution of multi-word units and discursive practices in business meetings (McCarthy and Handford 2004; Handford 2010). ABOT features texts of 'informal, unplanned workplace interactions between co-workers in office settings' (Koester 2010: 13), and these texts have been analysed using a discourse approach to corpus-based analysis in investigating

the performance of communicative functions in the workplace from speech acts (Koester 2002) and relational sequences ('transactional-plus-relational talk') to conversation analysis (Koester 2004). Although the dominant themes in these studies highlight localised or national settings (i.e. not with international business), various cultural groups may also be explored across variables, such as gender, speakers' first language background, register and genre overlap and specific task foci. And finally, workplaces with employees who use augmentative and alternative communication (AAC) devices and their non-AAC counterparts provided spoken data for the AAC and Non-AAC User Workplace Corpus (ANAWC) (Pickering and Bruce 2009). This highly specialised corpus represents machine-based language production, annotated for communicative items such as pauses and wait times, small talk markers (e.g. conversation openers) space fillers (Di Ferrante 2013), part-of-speech (POS) (tags and transitions/overlaps. Participants in eight target workplaces in the United States were given voice-activated recorders to be used for a full week of data collection, capturing a range of workplace events. The ANAWC broadly interprets the definition of office-based settings, and recordings range from IT offices to warehouse floors (Friginal et al. 2016).

Since the early 2000s, there has been an increasing number of outsourced call centre studies based on corpora conducted by researchers especially from the Philippines, India, Hong Kong and South Korea. Customer service call centre services in the United States and many 'English-speaking' countries (e.g. Australia, the United Kingdom and Canada) have been outsourced to overseas locations, primarily in order to lower operational costs incurred by maintaining these call centres locally. 'Outsourcing' is defined by the World Bank as 'the contracting of a service provider to completely manage, deliver and operate one or more of a client's functions (e.g. data centres, networks, desktop computing and software applications)' ('World Bank E-Commerce Development Report' 2003). From the mid-1990s, various call centre operations of several multinational corporations have transformed the nature of telephone-based customer services globally, and the expectations about the types of communication exchanges involved in these transactions (Beeler 2010; Friedman 2005; Friginal 2009a; Friginal and Cullom 2014). Satellite telecommunication and fibre-optic technologies have allowed corporations to move a part of their operations to countries such as India and the Philippines, which offer viable alternatives to the high cost of the maintenance of these call centres domestically. In the United States, 'third-party' call centres specialising in training and hiring Indian and Filipino customer service representatives (or 'agents') have increasingly staffed many U.S. companies for very low salaries by current standards (Friginal 2013a; Vashistha and Vashistha 2006).

Unlike in many other cross-cultural business and workplace settings, such as teleconferencing in multinational company meetings or negotiations in international commerce and trade, business communications in outsourced call centres have clearly defined roles, power structures and standards against which the satisfaction levels of customers during and after the transactions are often evaluated relative to agents' language and service performance (Cowie and Murty 2010; Forey 2016; Lockwood et al. 2009). Callers typically demand to be given the quality of service they expect or can ask to be transferred to an agent who will provide them the service they prefer. Offshore agents' use of language and explicit manifestations of pragmatic skills naturally are scrutinised closely when defining 'quality' during these outsourced call centre interactions. In contrast, for a foreign businessman in many intercultural business meetings, there may be limited pressure to perform following a specific (i.e. native speaker or L1) standard in language, as many business associates are often willing to accommodate linguistic variations and cultural differences of their counterparts in negotiations and performance of tasks, focusing, instead, on simply completing

the transaction successfully (Cowie 2016; Friginal 2009a, 2013b; Hayman 2010). These transactions in outsourced call centres, therefore, have produced a relatively new register of workplace discourse involving a range of variables not present in other globalised business or international and interpersonal communication settings.

In the United States and also many other developed countries, a significant number of healthcare communications have involved cross-cultural factors and contexts. Healthcare providers such as doctors and nurses have increasingly come from overseas locations. Specifically, foreign-educated nurses populate many hospitals in the United States and several European countries, handling a range of patient care services and are extensively trained and evaluated as to their language usage as well as their medical and patient care skills. Healthcare studies have primarily focused on doctor–patient and nurse–patient clinical interactions, patient narratives and documentation of consultations, often using qualitative methodologies – conversation analysis, ethnographies and interactional sociolinguistics (Frankel 1984; Mischler 1984; Staples 2016). These studies have provided valuable insight into the functional phases of healthcare interactions as well as the progression of discourse between patients and providers (Pickering et al. 2016).

Corpora and corpus-based approaches allow researchers to establish patterns of language used in healthcare settings, as well as to clearly describe discourse contexts and functional norms. Although corpus linguistic studies are still relatively rare in the study of healthcare communication, there is a growing number that have focused on multiple lexico-grammatical features within doctor–patient interactions and connect these linguistic features to more patient-centred communication styles. Drawing on previous work by Biber (1988), Thomas and Wilson (1996) illustrated how a doctor who was identified by patient reports as more patient-centred used more linguistic features associated with interactional involvement (e.g. pronouns, discourse markers, present tense). Cortes (2015) explores the linguistic features in patients' stories (i.e. patient narratives) about how they manage diabetes. Cortes discussed the various ways in which patients understand and experience their illness and how that relates to patient compliance. Hesson (2014) made use of corpora to investigate features of physician discourse in doctor–patient interactions across medical specialties (e.g. oncology vs. diabetes) and physician characteristics (e.g. sex, years in practice), focusing on the distributions of adjective complement clause types (finite vs. non-finite) as representative of stronger and weaker statements of importance. The study reported that physicians with more years in practice used stronger and specifically more direct statements, possibly due to earlier socialisation within a medical discourse community less focused on patient-centred care (Pickering et al. 2016).

Skelton and colleagues in the late 1990s also introduced methods of concordancing and collocational analysis to examine key phrases used to mitigate asymmetry within provider–patient interactions. Skelton and Hobbs (1999) showed how doctors use downtoners (e.g. *just, little*) when providing directives during physical exams. Modals and likelihood adverbs have been identified as a method of providing suggestions to patients and to discuss possible future states of affairs. A number of studies have also emphasised the importance of conditionals in medical encounters to perform some of these same functions (Adolphs et al. 2004; Ferguson 2001; Holmes and Major 2002; Skelton and Hobbs 1999; Skelton et al. 1999). Due to patient privacy and confidentiality issues restricting the use of corpora from hospital interactions, many of these studies do not share data which might enable further analyses (Pickering et al. 2016). Finally, Staples (2015, 2016) investigates the language of foreign-educated nurses, mostly from the Philippines, in interactions with patients from a U.S. hospital as compared with their U.S.-educated counterparts. She outlined an overall

structure of nurse–patient discourse, reporting that nurses make regular use of possibility modals, conditionals and first- and second-person pronouns to create a more patient-centred environment in interactions. The use of past tense, yes/no questions and backchannels to discuss patients’ psychosocial issues and of prediction modals to provide indications within the physical examination were consistently employed by nurses as they communicated with patients. Differences were found in the nurses’ communicative styles based on their background in terms of first language, country of origin and the country in which they received their education and training.

Sample analysis

This section provides a sample analysis of cross-cultural business interaction in outsourced call centres, making use of a corpus of recorded calls collected in the Philippines (with Filipino call-takers supporting American callers). Friginal’s corpus of Outsourced Call Center Interactions (Friginal 2009a) was provided by a U.S.-owned call centre company based in the Philippines and India. The corpus was carefully collected to represent various speaker categories obtained from Filipino and Indian agents and American callers. Situational sub-registers such as task categories (e.g. troubleshooting service calls vs. product purchase) or level of transactional pressure (e.g. high-pressure calls from complaint-oriented transactions vs. low-pressure calls from product questions or enquiries about service) were also coded. In addition to interactional studies that focus on speaker demographics, Friginal’s analyses of outsourced call centre interactions included discourse strategies such as stance and politeness markers, communicative strategies and miscommunication and issues of national and social identity. Arguably, the corpus approach represents the domain of outsourced call centres more extensively than studies based on only a few interactions of caller–agent discourse (Friginal 2009b). For example, correlational data between call centre agents’ patterns of speech, language ability and success or failure of transactions contribute valuable insights that could be used to improve the quality of training, and, consequently, of service. Generalisable information derived from such a representative corpus of call centre transactions can better inform and direct language training programmes and possibly support (or not) the viability of these outside of the United States.

For the sample analysis presented here, a test corpus of Filipino, Indian and American (i.e. agents based in the United States) call centre transactions ($N = 655$ total texts from two main types of tasks: troubleshooting and product enquiry/order) was used. The calls that qualified in this test corpus ranged from 5 to 25 minutes in duration and were obtained from Friginal’s dataset collected in 2009. Additional texts from U.S. and Indian agents were collected from 2010 to 2012. These transactions were transcribed into machine-readable text files following conventions used in the collection of the service encounter corpus the TOEFL 2000 Spoken and Written Academic Language (T2KSWAL). (See Biber 2006 for a description of this corpus.) Personal information about the callers, if any (e.g. names, addresses, phone numbers, etc.) was consistently and strictly replaced by different proper nouns or a series of numbers in the transcripts to maintain the interlocutors’ privacy.

Corpus-based multidimensional analysis (MDA) was used to describe the systematic patterning of linguistic features used by Filipino, Indian and American call centre agents as they serve their customers on the telephone. The MDA framework was developed by Biber (1988, 1995) in establishing the statistical co-occurrence of linguistic features from corpora. The concept of linguistic co-occurrence, which is the foundation of MDA, can be introduced by pointing out intuitively the common differences in the linguistic composition of

various types of registers. Specifically, for call centre agents, this framework allows for the examination of the agents use of the dimensional patterns of speech based on their cultural backgrounds and their transactional tasks (e.g. troubleshooting task vs. order placement/product enquiry). In sum, the primary research question pursued here is: Do potential cross-cultural differences influence agents' use of multidimensional features of telephone-based customer service talk?

The following comparisons illustrate how the three groups of agents made use of 'addressee-focused, polite and elaborated speech' vs. 'involved and simplified narrative'. These features (shown in Table 15.1) comprise a linguistic dimension that differentiates between the distributions of second-person pronouns (*you/your*), polite and elaborated discourse (e.g. politeness and respect markers, longer average length of turns, nouns and nominalisations) and involved (e.g. private verbs) and simplified narrative portraying how informational content is produced by agents in customer service transactions.

The primary result of this sample analysis is summarised in Figure 15.1, which shows the average dimension scores of Filipino, Indian and American agents along a positive and negative scale for 'addressee-focused, polite and elaborated speech' (positive side) vs. 'involved and simplified narrative' (negative side). These dimension scores reveal differences in the way the three groups of agents from two task groups make use of these co-occurring linguistic features. Filipino agents plot on the positive side of the scale, while Indian and American agents are both on the opposite, negative side. The use of addressee-focused, involved production features and politeness and respect markers characterises the overall nature of outsourced call centre transactions, but Filipino agents appear to use these features overwhelmingly compared to the two other agent groups. Service encounters commonly allocate for courteous language and the recognition of roles (e.g. server/servee), and call centre agents are expected to show respect and courtesy when assisting customers. Of interest here is the variation between these three groups who are, in fact, dealing with very similar contexts.

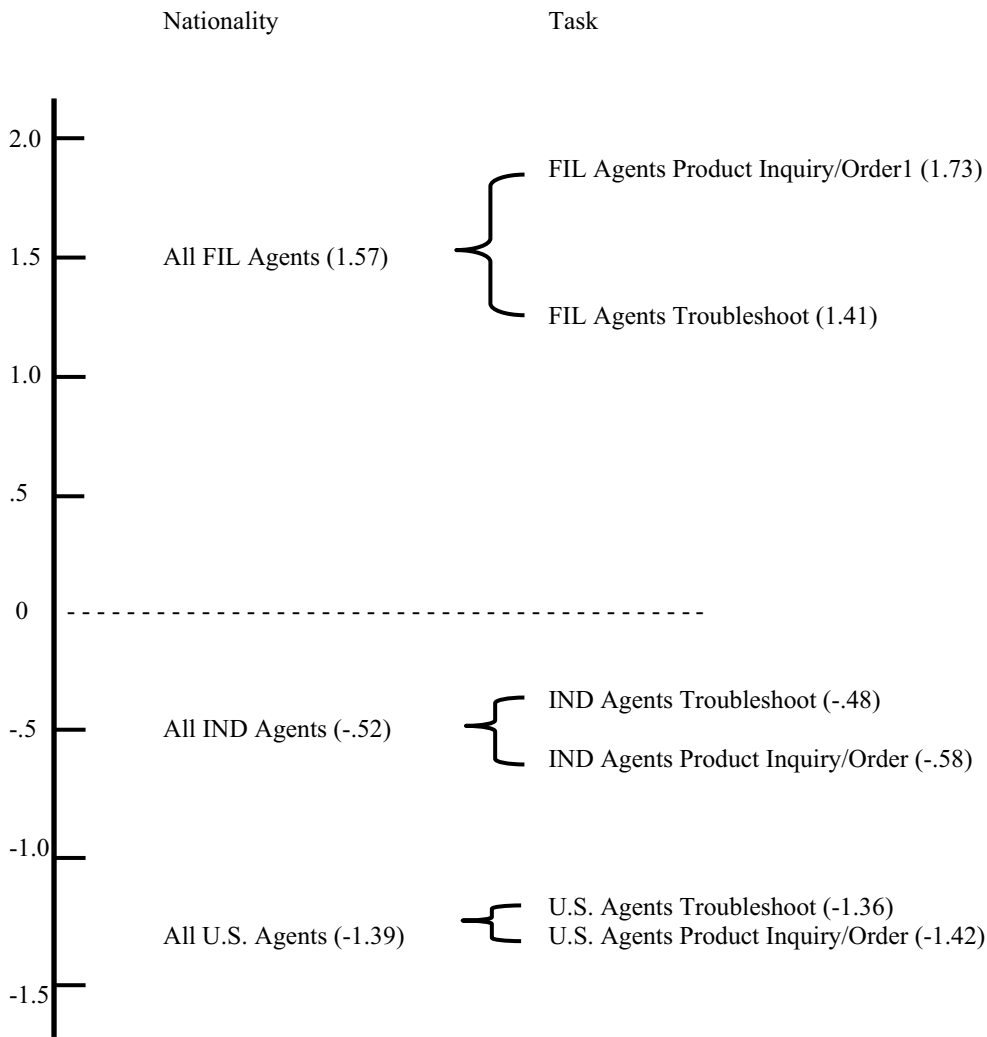
The task-based distinctions between the groups of agents in troubleshooting and enquiry/order transactions are interesting and worthy of more detailed analyses. The variation in the use of production features by groups of Filipino, Indian, and U.S. agents may be influenced by the nature of the issue or concern raised by the caller. Callers with complicated issues and those who are reporting dissatisfaction with service will require more elaborated responses and longer turns. Overall, it appears that Filipino agents prioritise friendliness and the maintenance of customer service persona more than directness and quick resolution of caller issues. As part of training, the Filipino agent in Text Sample 1 was probably coached to establish rapport with the caller by inserting an apology (e.g. *"I apologise for*

Table 15.1 Linguistic dimensions of outsourced call centre discourse

Dimension	Features
Addressee-Focused, Polite and Elaborated Speech vs. Involved and Simplified Narrative	Positive: Second-person pronouns, average word length, <i>please</i>, nouns, possibility modals, nominalisations, average length of turns, <i>thanks</i>, <i>ma'am/sir</i> Negative: Pronoun <i>it</i>, first-person pronouns, past tense verbs, <i>that</i> deletion, private verbs, WH clauses, perfect aspect verbs, <i>I mean/you know</i>, verb <i>do</i>

Source: Original 2009a, 2013b

ADDRESSEE-FOCUSED, POLITE, AND ELABORATED SPEECH



INVOLVED AND SIMPLIFIED NARRATIVE

Figure 15.1 Comparison of average agent dimension scores for ‘addressee-focused, polite and elaborated speech’ vs. ‘involved and simplified narrative’

the inconvenience sir, I'll, let me explain on that ok?") and by providing additional details as she assisted the caller in checking the remaining balance on the phone card (e.g. “*For the meantime Mr. [name], you can also check your balance on your phone by calling 1-800-000-0000, and that is a free call always.*”).

The comparisons in Figure 15.1 show that Filipino agents had the greatest number of politeness markers *ma'am/sir, please, thanks* and all markers of discourse elaboration. Indian agents have shorter turns than Filipinos and the transactions maintain direct question–answer

sequences. American agents made use of very few politeness and elaboration features, but they use the greatest number of private verbs and transition markers (*I mean/you know*). Texts Samples 1, 2 and 3 illustrate how the agent groups performing similar tasks (Product Enquiry/Order) differ in their use of the features of this dimension.

Text Sample 1. Call excerpt from FIL Product Enquiry/Order (Friginal 2013b)

- Agent:* **Thank you** for calling [Phone Company] **Payment Services**, my name is [agent name], how **can** I help **you**?
- Caller:* *Yes, uh, when are you guys gonna go back telling us when how much time is left on these phone cards? I mean on these phones?*
- Agent:* **I apologize** for the **inconvenience** **sir**, I'll, let me explain on that ok? **Please**, give me your **cell phone number** so I can check on **your minutes**.
- Caller:* *[cell phone number], I think it has run out because I wanted to use it but it said it didn't have enough time.*
- Agent:* Ok, let me just verify the charges at the **moment**, **please** give me **your name** and **address** on the **account please**
- Caller:* *[caller name and address]*
- Agent:* **Thank you** for that Mr. [caller name], let me just pull out **your account** to check **your balance**, ok, Sir? Mr. [caller name], **you** have now zero **balance** on the **account** and uh, ok Mr. [name], **you** are notified of **your balance** when **you** reached below **\$10**, below
- Caller:* *There never was a word said anytime. I never heard anything. How am I supposed to be notified?*
- Agent:* I see, well **sir** do **you**, ok just a **moment**, while I check on **your account**. Ok, **sir** did **you** give out any **e-mail address** where we can send **updates** regarding **your account**?
- Caller:* *Yeah I did, but I don't know, my computer is down right now.*
- Agent:* For the meantime Mr. [name], **you can** also check **your balance** on **your phone** by calling 1-800-000-0000, and that is a free **call** always. Just choose the **option** for **you** to receive the **minutes** on **your** account, either through uh, or via **text message**.

Text Sample 2. Call excerpt from IND Product Enquiry/Order (Friginal 2013b)

- Caller:* *I wanted to check, you could, could you help me load minutes into this [brand name] phone? I believe you have my account information? My husband set it up for us. What do you need?*
- Agent:* I just need the **phone number, please**.
- Caller:* *[caller provides phone number] and it's under my name*
- Agent:* Could **you** verify **your** name for me **please**?
- Caller:* *[caller provided name]*
- Agent:* Ok **thank you** for that. **You** still have a total of a total of, uh, \$12. How much do **you** need to add? I noticed that there is also a \$10 credit that **you** have not activated, I can activate that for **you** if **you** want.

Caller: Sure. Please add \$20. Could you use my card on record?
Agent: Yes Ms. [name]
Caller: Say that again?
Agent: Yes, sure, **I mean I can ma'am.**
Caller: Thanks.

Text Sample 3. Call excerpt from U.S. Product Enquiry/Order

Agent: Ok **thank you** for that. **You** still have a total of a total of, uh, \$12. How much do **you** need to add? I noticed that there is also a \$10 credit that **you** have not activated, I can activate that for **you** if **you** want.
Caller: Sure. Please add \$20. Could you use my card on record?
Agent: Yes Ms. [name]
Caller: Say that again?
Agent: Yes, sure, **I mean I can**
Caller: Thanks.

Clearly, the corpus approach to cross-cultural communication has multiple applications in further understanding the complexities of this communicative domain. Frequency data provide researchers with the ability to run statistics and related tests that may lead to concrete conclusions and accurate generalisations. However, there are also clear limitations and areas for improvement, especially in the broader areas of research in digital humanities. For example, very important linguistic components such as the segmental and supra-segmental features of oral discourse are not easily captured in transcribed texts. Even with annotated corpora such as the HKCSE or Friginal's corpus of outsourced call centre interactions, the corpus approach is only able to account for variation in texts. Oral discourse features, very important in further interpreting speaker cross-cultural sentiments, feelings and attitudes will have to be explored using supplemental approaches and technologies. Recently, there have been promising advancements in speech recognition devices that allow for immediate sentiment analysis. The use of spectrographic displays in coding pronunciation and intonation of utterances has also been explored.

Digital photography and cross-cultural communication

Background

Digital recording and photography have been used as a tool in research methodologies related to language and culture for many years. The evolution of visual technologies has led to more advanced methodological innovations that enhance the understanding of communication and how it functions across settings or contexts. For example, the use of digital recording and photography has emerged as a primary method for investigating the *linguistic landscape* (Landry and Bourhis 1997) of spaces, particularly public spaces. The investigation of linguistic landscape (LL) focuses on language use and perceptions of language in public spheres, including cross-cultural contexts, and facilitates the understanding of communication and how it functions across various multilingual and multicultural settings.

Digital photography and linguistic landscape research

The concept of LL has been employed by scholars over the past few decades for several purposes and in various spaces and locations. One of the early and most commonly employed description of LL as a research approach is the use of

the language of public road signs, advertising billboards, street names, place names, commercial shop signs, and public signs on government buildings combines to form linguistic landscape of a given territory, region, or urban agglomeration.

(Landry and Bourhis 1997: 25)

Generally, the study of LL is a useful tool to facilitate the understanding of the linguistic diversity of a particular space and location, how language is used and by whom in multilingual and multicultural settings (Ben-Rafael et al. 2006) and as supporting input for language learning (Cenoz and Gorter 2006; Sayer 2009; Malinowski 2015). LL studies may focus on particular types of signs such as shops' signs and names (Dimova 2007; MacGregor 2003), while others may use main shopping streets and business domains and could include all visible or displayed texts (Cenoz and Gorter 2006).

The study and conceptualisation of LL has also been evolving over the years, especially with advancements in digital audio and video technology. While Landry and Bourhis's definition seems to focus on 'signs,' many scholars are now expanding their operationalisation of LL to include broader areas of language used in the public spaces, for example:

language in the environment, words and images displayed and exposed in public spaces, that is the center of attention in this rapidly growing area.

(Shohamy and Gorter 2009)

More recently scholars have described LL to include 'verbal texts, images, objects, placement in time and space as well as human beings' (Shohamy and Waksman 2009: 314). The description and analysis of LL can expand perspectives about language use in and across contexts, in particular multicultural and multilingual contexts. For example, Landry and Bourhis (1997) use LL to inform their studies and analyse 'the most observable and immediate index of relative power and status of the linguistic communities inhabiting a given territory' (p. 29). In this sense, observations of the LL provide a method of empirical investigation into the presence of specific languages or linguistic communities in a given area signalling the sociolinguistic composition in a territory and uncovering the power and status of languages and the purposes they serve in public. LL uncovers the social realities of a designated space (Ben-Rafael et al. 2006) and represents 'observable manifestations of circulating ideas about multilingualism' (Shohamy 2006: 110).

Digital photography and cross-cultural linguistic landscape studies

Language awareness and LL

Daganais and colleagues (2009) utilise LL data in a Canadian context to research children's perceptions of the texts, multilingualism and language diversity in their communities, including the interactions of social actors from diverse communities. The participants in the study included students in grade 5 and their teachers and parents in two schools

in Vancouver (French immersion programme) and Montreal (French school district). The study was multipart and longitudinal. After gathering data in the form of fixed signs from zones near the two schools, researchers guided students in analysing signs and then collecting more data. The initial signs that were collected represented the diverse range of language use in the neighbourhoods and include unilingual and bilingual signs in French and English, as well as multilingual signs using languages other than French and English. Initially the students participated in a focus group regarding language in their communities and were then videotaped doing language awareness activities over a period of four months. Focus groups were held after the four-month period. Teachers and parents were also interviewed focusing on children's responses and their language contact.

Some representative examples of the activities utilised in Montreal include having students describe their neighbourhoods and the languages of people around them. Students also created murals utilising drawings of places in their neighbourhoods along with short texts related to the drawings. During the following days of the study, students were guided to reflect and analyse their own and others' drawings. Researchers found that students themselves were not representative of the linguistic diversity in their neighbourhoods, although they were aware of the languages they heard as well as the diverse populations.

Similar activities were conducted with the students at the Vancouver school, but with hand-drawn maps to guide them through the community, complementing a mapping lesson in their social studies class. Dagenais and colleagues drew on the work of Scollon and Scollon (2003) to support the idea that LL serves as a tool to develop the critical literacy of students by guiding them through three levels of analysis related to what they observed, including 'the geographical location, sociological importance and linguistic function of media in the landscape' (p. 45) and later helping them to reflect on the human geography and history of place in their specific multicultural context. This study demonstrates that observing the LL in language awareness activities is beneficial in facilitating lessons through a critical lens regarding language diversity and students' varying literacies. Concluding remarks by the researchers called for an expansion of pedagogical approaches related to diversity 'beyond the typical focus on religion, culture and ethnicity explore din liberal multiculturalism or intercultural education' (p. 266).

Second language acquisition and communicative pragmatics in LL

Cenoz and Gorter (2006) highlight the role of digital photography and LL in second language acquisition as a source of input. Their major assertion is that LL for language learners can act in informal learning contexts as authentic input. In their view, items in the LL can be considered actual speech acts, and exposure to these could increase learners' pragmatic competence, which is one of the most important aspects of communicative competence (Bachman 1990; Celce-Murcia et al. 1995). From the perspective of learning languages for cross-cultural communication, 'authentic input' raises learners' awareness of what is actually required to communicate successfully and pragmatically in a particular context. Cenoz (2007) notes that in intercultural communication, pragmatic failure may not be recognised by those participating in an exchange, and the interlocutor may be perceived as impolite or uncooperative. In sum, 'the LL opens the possibilities of having access to authentic input that can raise learners' awareness about the realization of different speech acts' (Cenoz and Gorter 2006: 271).

In Mexico, Sayer (2009) employs linguistic landscape as a pedagogical resource for his students learning English in a foreign language context as a way of connecting

in-class lessons to real-world language use. He guides his students in analysing English and its social meaning and significance in the linguistic landscape of Oaxaca. He finds that English serves communicative purposes both across cultures (for visiting tourists) and also intraculturally. Intraculturally speaking, English functions socially in various ways, for example, to portray and promote items as sophisticated in marketing technology, or fashionable and modern in the fashion industry, and also to project political identities through social protest (Freire 1970). Sayer notes that English in Oaxaca is increasingly globalised, and, as noted by Pennycook (2007), it exemplifies the language as a transnational cultural product. According to Sayer, in the Oaxacan context English has acquired local meanings as individuals learn it and use it for their own purposes (2009). Sayer's framework for using LL in the classroom has students taking on the role of 'language detectives' and photographers. He suggests having students collect their own data via digital photography and guiding students to come up with categories for understanding the meaning and function of language in whatever context they are in and having them present and defend their analyses and interpretations of their photographs. Besides contextualising English language learning beyond the classroom, using LL as a pedagogical tool facilitates understanding of how language is used across societies and in the students' own sociolinguistic contexts.

While the aforementioned studies offer insights in using the LL as a way to enhance and contextualise the acquisition of linguistic systems and pragmatic language skills, other studies have focused more on the critical aspects of using LL in learning to promote critical awareness and even activism. Shohamy and Waksman (2009) argue for participatory and inquiry-based educational activities across multilingual contexts in the interest of critical pedagogy, activism and language rights. For example, one study of a South Korean undergraduate 'Introduction to Intercultural Studies' class had undergraduates in a translation and interpretation programme conduct small research studies in order to develop an understanding of language and place and to more critically understand symbolic and figurative applications of language across contexts.

Malinowski (2015) offers a conceptual framework for providing a link between recent developments in LL theory and methods and language and literacy learning. He proposes the concept of spatialising learning based on the principle of *thirdness*. His framework invokes Kramsch's (2011) concept of 'third place' or:

The capacity to recognize the historical context of utterances and their intertextualities, to question established categories like German, American, man, woman, White, Black and place them in their historical and subjective contexts. But it is also the ability to resignify them, reframe them, re-and trans-contextualize them and to play with the tension between text and context.

(Kramsch 2011: 359)

The point here is to make the teaching and learning experience one that allows students and teacher go beyond 'communicative and cultural binaries' (Malinowski 2015: 101). Malinowski's view builds on Trumper-Hecht's (2010) and other research that reinterprets Lefebvre's (1991) conceptualisation of space being conceived, perceived and lived in the language classroom. Using digital photography for learning about language and culture would go beyond taking photos of and noting target language signage (perceived space) and beyond talking to local actors about their perceptions of the signage (lived space). According to Malinowski, adopting the idea of a third space, students

would do both of these and more, also comparing their own experiences and those of others to outside perspectives on local realities in readings, films, maps, and other texts *conceived space*)[. . .] With language learners striving to interpret the linguistic, social, and political significance of multilingual signs in oft-contested spaces, the linguistic landscape as concept, field, and place may yet find itself transformed through the conscious incorporation of pedagogical insights and practices.

(2015: 109)

This conceptualisation of LL as a pedagogical tool incorporates critical thought and awareness in the teaching and learning experience and empowers students to think across cultures.

In the context of digital humanities, analysis of the LL provides a structure for analysing public texts as cultural artefacts. The potential of utilising the aforementioned methods, digitising the LL and creating digital or interactive experiences using these landscapes is promising and would lend itself to making available cross-cultural experiences to those who may not be able to immerse themselves in another culture, but may have access to a computer and thus exposure to a more complete contextual understanding that they would otherwise not have access to.

Further reading

- 1 Belcher, D. and Nelson, G. (eds.) (2013). *Critical and corpus-based approaches to intercultural rhetoric*. Ann Arbor, MI: University of Michigan Press.
This collection of corpus-based studies in intercultural communication explores critical and frequency-based perspectives to explain what is meant by 'culture' and how this affects research and teaching along descriptions of new forms of literacy. Most chapters highlight the contribution of corpus linguistic tools which, according to the editors, have greatly impacted how intercultural rhetoric is researched. The volume has four parts: *Corpus and Critical Perspectives*, *Critical-Analytical Approaches*, *Corpus-Based Approaches*, and *Next Steps*.
- 2 Friginal, E. (2009). *The language of outsourced call centers*. Amsterdam: John Benjamins.
Friginal explores a large-scale corpus representing the typical kinds of interactions and communicative tasks in outsourced call centres located in the Philippines and serving American customers. The specific goals of this book are to conduct a corpus-based register comparison between outsourced call centre interactions, face-to-face American conversations and spontaneous telephone exchanges and to study the dynamics of cross-cultural communication between Filipino call centre agents and American callers, as well as other demographic groups of participants in outsourced call centre transactions (e.g. gender of speakers, agents experience and performance and types of transactional tasks).
- 3 Hélot, C., Barni, M. and Janssens, R. (2012). *Linguistic landscapes, multilingualism and social change*. Frankfurt: Peter Lang.
This book offers insights and perspectives on the analysis and interpretation of linguistic landscape across various social and cultural contexts and its potential for reflecting social change. It is a collection of presentations from the 3rd International Linguistic Landscape workshop held in 2010 at the University of Strasbourg.
- 4 Pickering, L., Friginal, E. and Staples, S. (eds.) (2016). *Talking at work: Corpus-based explorations of workplace discourse*. London: Palgrave Macmillan.
This book offers original corpus research in a range of workplace contexts, including office-based settings, call centre interactions and healthcare communication. Chapters in this edited volume bring together leading scholars in the field of corpus analysis in workplace discourse and include data from multiple corpora. Employing a range of qualitative and quantitative analytic approaches,

including conversation analysis, linguistic profiling and register analysis, the book introduces unique specialised corpus data in the areas of augmentative and alternative communication, nursing and cross-cultural communication, among others.

References

- Adolphs, S., Brown, B., Carter, R., Crawford, P. and Sahota, O. (2004). Applying corpus linguistics in a health care context. *Journal of Applied Linguistics* 1(1): 9–28.
- Anthony, L. (2014). *AntConc (Version 3.3.5). [Software]*. Available at: www.antlab.sci.waseda.ac.jp/
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Beeler, C. (2010, August 25). *Outsourced call centers return to US homes*. NPR Online. Available at: www.npr.org/templates/story/story.php?storyId=129406588&ps=cprs (Accessed 27 August 2010).
- Belcher, D. and Nelson, G. (eds.) (2013). *Critical and corpus-based approaches to intercultural rhetoric*. Ann Arbor, MI: University of Michigan Press, pp. 127–153.
- Ben-Rafael, E., Shohamy, E., Hasan Amara, M. and Trumper-Hecht, N. (2006). Linguistic landscape as symbolic construction of the public space: The case of Israel. *International Journal of Multilingualism* 3(1): 7–30.
- Biber, D. (1988). *Dimensions of register variation*. Cambridge: Cambridge University Press.
- Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge, UK: Cambridge University Press.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
- Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press.
- Biber, D., Reppen, R. and Friginal, E. (2010). Research in corpus linguistics. In R.B. Kaplan (ed.), *The Oxford handbook of applied linguistics* (2nd ed.). Oxford: Oxford University Press, pp. 548–570.
- Brazil, D. (1997). *The communicative value of intonation in English*. Cambridge: Cambridge University Press.
- Celce-Murcia, M., Dornyei, Z. and Thurrell, S. (1995). Communicative competence: A pedagogically motivated framework with content specifications. *Issues in Applied Linguistics* 6: 5–35.
- Cenoz, J. (2007). The acquisition of pragmatic competence and multilingualism in foreign language contexts. In E. Alcon and M.P. Safont (eds.), *Intercultural language use and language learning*. Amsterdam: Springer, pp. 223–244.
- Cenoz, J. and Gorter, D. (2006). The linguistic landscape as an additional source of input in second language acquisition. *IRAL – International Review of Applied Linguistics in Language Teaching* 46(3): 267–287.
- Cheng, W., Greaves, C. and Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics* 11(4): 411–433.
- Cheng, W., Greaves, C. and Warren, M. (2008). *A corpus-driven study of discourse intonation*. Amsterdam: John Benjamins.
- Cheng, W. and Warren, M. (2006). I would say be very careful of . . . : Opine markers in an intercultural business corpus of spoken English. In J. Bamford and M. Bondi (eds.), *Managing interaction in professional discourse: Intercultural and interdiscoursal perspectives*. Rome: Officina Edizioni, pp. 46–57.
- Connor, U. (1996). *Contrastive rhetoric: Cross-cultural aspects of second language writing*. Cambridge, U.K.: Cambridge University Press.
- Cortés, V. (2015). Using corpus-based analytical methods to study patient talk. In M. Antón and E. Goering (eds.), *Understanding patients' voices*. Amsterdam: John Benjamins, pp. 23–49.
- Cowie, C. (2016). *Imitation and accommodation in international service encounters: IT workers and call centres in India*. Paper presented at Sociolinguistics Symposium 21, Murcia, Spain, 17 June.

- Cowie, C. and Murty, L. (2010). Researching and understanding accent shifts in Indian call center agents. In G. Forey and J. Lockwood (eds.), *Globalization, communication and the workplace*. London: Continuum, pp. 125–146.
- Dagenais, D., Moore, D., Sabatier, C., Lamarre, P. and Armand, F. (2009). Linguistic landscape and language awareness. In E. Shohamy and D. Gorter (eds.), *Linguistic landscape: Expanding the scenery*. New York, NY: Routledge, pp. 253–269.
- Di Ferrante, L. (2013). *Small talk at work: A corpus-based discourse analysis of AAC and Non-AAC device users*. Unpublished PhD dissertation. Texas A&M University-Commerce, Commerce, TX.
- Dimova, S. (2007). English shop signs in Macedonia. *English Today* 23(3–4): 18–24.
- Ferguson, G. (2001). If you pop over there: A corpus-based study of conditionals in medical discourse. *English for Specific Purposes* 20: 61–82.
- Forey, G. (2016). *Prestige, power and empowerment for workers in the call centre context*. Paper presented at Sociolinguistics Symposium 21, Murcia, Spain, 17 June.
- Frankel, R.M. (1984). From sentence to sequence: Understanding the medical encounter from micro-interactional analysis. *Discourse Processes* 7: 135–170.
- Freire, P. (1970). *Pedagogy of the oppressed*. New York: Seabury Press.
- Friedman, T.L. (2005). *The world is flat: A brief history of the twenty-first century*. New York: Picador.
- Friginal, E. (2009a). Threats to the sustainability of the outsourced call center industry in the Philippines: Implications to language policy. *Language Policy* 5(1): 341–366.
- Friginal, E. (2009b). *The language of outsourced call centers: A corpus-based study of cross-cultural interaction*. Philadelphia: John Benjamins Publishing Company.
- Friginal, E. (2013a). 25 years of Biber’s multi-dimensional analysis: Introduction to the special issue. *Corpora* 8(2): 137–152.
- Friginal, E. (2013b). Linguistic characteristics of intercultural call center interactions: A multi-dimensional analysis. In D. Belcher and G. Nelson (eds.), *Critical and corpus-based approaches to intercultural rhetoric*. Ann Arbor, MI: University of Michigan Press, pp. 127–153.
- Friginal, E. and Cullom, M. (2014). Saying ‘No’ in Philippine-based outsourced call center interactions. *Asian Englishes*. doi:10.1080/13488678.2013.872362
- Friginal, E. and Hardy, J.A. (2014). *Corpus-based sociolinguistics: A guide for students*. New York: Routledge.
- Friginal, E., Lee, J., Polat, B. and Roberson, A. (2017). *Exploring spoken English learner language using corpora: Learner talk*. London: Palgrave Macmillan.
- Friginal, E., Pickering, L. and Bruce, C. (2016). Narrative and informational dimensions of AAC discourse in the workplace. In L. Pickering, E. Friginal and S. Staples (eds.), *Talking at work: Corpus-based explorations of workplace discourse*. London: Palgrave Macmillan, pp. 27–54.
- Friginal and Staples. (2016). cited in text but does not appear in reference list.
- Handford, M. (2010). *The language of business meetings*. Tokyo: Cambridge University Press.
- Hayman, J. (2010). Talking about talking: Comparing the approaches of intercultural trainers and language teachers. In G. Forey and J. Lockwood (eds.), *Globalization, communication and the workplace*. London: Continuum, pp. 147–158.
- Hesson, A. (2014). Medically speaking: Mandative adjective extraposition in physician speech. *Journal of Sociolinguistics* 18(3): 289–318.
- Hofstede, G. (2001). *Culture’s consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage.
- Holmes, J. (2006). Sharing a laugh: Pragmatic aspects of humor and gender in the workplace. *Journal of Pragmatics* 38: 26–50.
- Holmes, J. and Major, G. (2002). Nurses communicating on the ward: The human face of hospitals. *Kai Tiaki, Nursing New Zealand* 8(11): 4–16.
- Kirschenbaum, M. (2014). What is ‘Digital Humanities,’ and why are they saying such terrible things about it? *Differences* 25(1): 46–63.
- Koester, A. (2002). The performance of speech acts in workplace conversations and the teaching of communicative functions. *System* 30(2): 167–184.

- Koester, A. (2004). Relational sequences in workplace genres. *Journal of Pragmatics* 36: 1405–1428.
- Koester, A. (2010). *Workplace discourse*. London: Continuum.
- Kramsch, C. (2011). The symbolic dimensions of the intercultural. *Language Teaching* 44(3): 354–367.
- Landry, R. and Bourhis, R.Y. (1997). Linguistic landscape and ethnolinguistic vitality. *Journal of Language and Social Psychology* 16(1): 23–49.
- Lefebvre, H. (1991). *The production of space*. Oxford: Blackwell.
- Lockwood, J., Forey, G. and Elias, N. (2009). Call center communication: Measurement processes in non-English speaking contexts. In D. Belcher (ed.), *English for specific purposes in theory and practice*. Michigan: University of Michigan Press, pp. 143–164.
- MacGregor, L. (2003). The language of shop signs in Tokyo. *English Today* 19(1): 18–23.
- Malinowski, D. (2015). Opening spaces of learning in the linguistic landscape. *Linguistic Landscape* 1(1–2): 95–113.
- Marra, M. (2012). English in the workplace. In B. Paltridge and S. Starfield (eds.), *The handbook of English for specific purposes*. Chichester, UK: John Wiley & Sons, Ltd, pp. 67–99.
- McCarthy, M. and Handford, M. (2004). Invisible to us: A preliminary corpus-based study of spoken business English. In U. Connor and T.A. Upton (eds.), *Discourse in the professions: Perspectives from corpus linguistics*. Amsterdam: John Benjamins, pp. 167–201.
- McEnery, T., Xiao, R. and Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. New York, NY: Routledge.
- Mischler, E. (1984). *The discourse of medicine*. Norwood, NJ: Ablex.
- Pennycook, A. (2007). *Global Englishes and transcultural flows*. London, UK: Routledge.
- Pickering, L. and Bruce, C. (2009). *AAC and Non-AAC Workplace Corpus (ANAWC)*. Atlanta, GA: Georgia State University.
- Pickering, L., Friginal, E. and Staples, S. (eds.) (2016). *Talking at work: Corpus-based explorations of workplace discourse*. London: Palgrave Macmillan.
- Sayer, P. (2009). Using the linguistic landscape as a pedagogical resource. *ELT Journal* 64(2): 143–154.
- Scollon, R. and Scollon, S.B.K. (2003). *Discourses in place: Language in the material world*. London: Routledge.
- Scott, M. (2010). *Wordsmith: Software tools for windows*. Available at: www.lexically.net/downloads/version6/wordsmith6.pdf.
- Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches*. New York: Routledge.
- Shohamy, E. and Gorter, D. (2009). *Linguistic landscape: Expanding the scenery*. London: Routledge.
- Shohamy, E. and Waksman, S. (2009). Linguistic landscape as an ecological arena: Modalities, meanings, negotiations, education. In E. Shohamy and D. Gorter (eds.), *Linguistic landscape: Expanding the scenery*. New York: Routledge, pp. 313–330.
- Sinclair, J. (2005). Corpus and text – basic principles. In M. Wynne (ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1–16). Oxford: Oxbow Books.
- Skelton, J.R. and Hobbs, F.D.R. (1999). Concordancing: Use of language-based research in medical communication. *The Lancet* 353: 108–111.
- Skelton, J.R., Murray, J. and Hobbs, F.D.R. (1999). Imprecision in medical communication: Study of a doctor talking to patients with serious illness. *Journal of the Royal Society of Medicine* 92: 620–625.
- Staples, S. (2015). *The discourse of nurse-patient interactions: contrasting the communicative styles of U.S. and international nurses*. Philadelphia: John Benjamins.
- Staples, S. (2016). Identifying linguistic features of medical interactions: A register analysis. In L. Pickering, E. Friginal and S. Staples (eds.), *Talking at work: Corpus-based explorations of workplace discourse*. London: Palgrave Macmillan, pp. 179–208.
- Stubbe, M., Lane, C., Hilder, J., Vine, E., Vine, B., Marra, M., Homes, J. and Weatherall, A. (2003). Multiple discourse analyses of a workplace interaction. *Discourse Studies* 5(3): 351–388.
- Tannen, D. (2000). ‘Don’t Just Sit There-Interrupt!’ Pacing and pausing in conversational style. *American Speech* 75(4): 393–395.
- Terras, M. (2016). A decade in digital humanities. *Journal of Siberian Federal University. Humanities & Social Sciences* 9: 1637–1650.

- Thomas, J. and Wilson, A. (1996). Methodologies for studying a corpus of doctor – patient interaction. In J. Thomas and M. Short (eds.), *Using corpora for language research: Studies in honour of Geoffrey Leech*. White Plains, New York: Longman, pp. 92–109.
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Philadelphia, PA: John Benjamins.
- Trumper-Hecht, N. (2010). Linguistic landscape in mixed cities in Israel from the perspective of ‘walkers’: The case of Arabic. In E. Shohamy, E. Ben-Rafael and M. Barni (eds.), *Linguistic landscape in the city*. Bristol: Multilingual Matters, pp. 235–251.
- Vashistha, A. and Vashistha, A. (2006). *The offshore nation: Strategies for success in global outsourcing and offshoring*. New York: McGraw-Hill.
- Vine, B. (2016). Pragmatic markers at work in New Zealand. In L. Pickering, E. Friginal and S. Staples (eds.), *Talking at work: Corpus-based explorations of workplace discourse*. London: Palgrave Macmillan, pp. 1–26.
- Vine, B. (2017). *Just and actually* at work in New Zealand. In E. Friginal (ed.), *Studies in corpus-based sociolinguistics*. New York: Routledge, pp. 197–218.
- Warren, M. (2004). //so what have YOU been WORKing on REcently//: Compiling a specialized corpus of spoken business English. In U. Connor and T. Upton (eds.), *Discourse and the professions: Perspectives from corpus linguistics*. Amsterdam: John Benjamins, pp. 115–140.
- Warren, M. (2011). Realisations of intertextuality, interdiscursivity and hybridisation in the discourses of professionals. In G. Garzone and M. Gotti (eds.), *Discourse, communication and the enterprise: Genres and trends*. Frankfurt: Peter Lang, pp. 91–110.
- Warren, M. (2013). ‘Just spoke to . . .’: The types and directionality of intertextuality in professional discourse. *English for Specific Purposes* 32(1): 12–24.
- Warren, M. (2014). ‘yeah so that’s why I ask you to er check’: (Im)politeness strategies when disagreeing. In S. Ruhi and Y. Aksan (eds.), *Perspectives on (Im)politeness: Issues and methodologies*. Newcastle: Cambridge Scholars Publishing, pp. 106–121.
- ‘World Bank E-Commerce Development Report’ (2003). World Bank Publications. Available at: <http://publications.worldbank.org/> (Accessed 12 January 2007).

Sociolinguistics

Lars Hinrichs and Axel Bohmann

Introduction

The digital turn has affected sociolinguistics in important ways: (1) it has facilitated research on traditional data types, enabling convenient and nearly loss-free storage, as well as management of written and audio-visual data; (2) it has brought about new data types that are relevant to sociolinguists, such as linguistic interactions conducted through digital media, or digitally administered experiments and surveys; and (3) it has led to improvements of research methods by orders of magnitude, allowing for the computational analysis of textual as well as audio-visual data. In this chapter we trace the impact of ‘the digital’ upon the sociolinguistic study of the English language at those three levels. Quantitative as well as qualitative, discourse-based and critical approaches will be considered. We then briefly present findings from a recent study that is digital at all three levels: a statistical analysis of language use based on a large corpus of digital texts that establishes multidimensional patterns of variation that stratify the data in terms of geographical space as well as genre. The conclusion returns to the discussion of the extent and quality of links among the fields of sociolinguistics and the digital humanities.

Definitions

Sociolinguistics is the study of language in social context. Its main interest lies in the ways in which human relationships, at both the small and the very large scale, have an impact on how language is realised. The main theoretical premise of sociolinguistics is that language should not simply be studied as a set of rules that operate independently of most situational influences, but instead as a rule-governed system whose output is subject, in regular ways, to situational and societal influences (Weinreich et al. 1968). The primary object of study in sociolinguistics, therefore, is linguistic variation. As such, sociolinguistics is to be seen as an intellectual reaction to the rise in the 1960s of Chomsky’s (e.g. 1965) philosophical view of language as a universal, genetically pre-determined capacity of the human brain, due to which the study of language ought to uncover the commonalities among human language systems, rather than explain variation among individual performances.

Nearly all sociolinguistic research, then, focuses on some aspect of linguistic variation, and yet the field is surprisingly broad at the level of methodological traditions. Arguably lying at its core, there is the quantitative tradition of *variationist sociolinguistics*, which looks to William Labov as its first major protagonist. This line of sociolinguistic work relies on surveys among larger samples of members of a speech community, defined primarily by geographic location. For instance, Labov's (1966) foundational study addressed language in New York City as a whole, and in some cases the Lower East Side. Survey data are interpreted statistically, with aspects of participants' linguistic performance, and variation therein, being correlated with independent variables such as social class, race, gender or age of the speaker, or the phonetic or grammatical context of the variable in question. Other types of sociolinguistic work take a qualitative approach to linguistic data: considering not only participant interviews but also ethnographic observation, public and private writing, cultural performances and artefacts, etc., as data, the social functions of linguistic forms and variation among them are explored. Some examples of work in this tradition are Heller (2010), Johnstone (2013), and Jaffe (1999).

Reflecting this methodological breadth, sociolinguists are hired by various types of academic departments. Some work in anthropology departments and conduct research in the tradition of linguistic anthropology, others in linguistics departments and yet others in language departments (e.g. English, Spanish or Middle Eastern Studies departments).

Among scholars working on English language and linguistics with a sociolinguistic interest, there is a subset invested in the study of world Englishes (or of global Englishes, or of varieties of English around the world). With an overarching interest in English as a language that has been and is being globally dispersed through, first, British colonialism and more recently modern-day globalisation (i.e. in 'English as a World Language'), these researchers take an interest in variation among different geographic and social varieties of English. Their research goals are centred on the description and explanation of similarities and differences among varieties of English around the world, and the methods and theories of both quantitative and qualitative sociolinguistics are regularly applied in this pursuit. Examples of quantitative-sociolinguistic work in world English studies include Schreier (2008) on variation in St. Helena English, and Patrick (1999) on variation in mesolectal Jamaican Creole. Qualitative examples include Pennycook (2003), a discourse analysis of the uses of English in rap lyrics from a Japanese act, and Wiebesiek et al. (2011), a study of attitudes toward South African English.

The fact that the field of sociolinguistics has grown since the second half of the twentieth century primarily out of studies of English-speaking, Western locales is of a certain practical convenience to researchers of the linguistics of English: large standardised corpora¹ and language-specific software tools are just two examples of resources that are less available for languages other than English. However, it has also created, as Smakman (2015) explains, a Westernising drift in sociolinguistic research around the world, even when the community under analysis is neither English-speaking nor Western. All of the most frequently used introductory textbooks to sociolinguistics are from Anglo-centric and/or Western locations, Smakman notes (p. 20), and a tally of 26 influential sociolinguists shows that all of them are from Western locales, and 19 out of 26 (73.1 percent) from the core of the English-speaking world: the United States, the UK, and Canada (p. 21). In studies of non-Western locales, theoretical assumptions of canonical sociolinguistics about the links between social factors and language use are often left unexamined. At the very least, this results in an incomplete or skewed view of non-Western locales. With sociolinguistic studies of non-Western locales being greatly outnumbered by those of the Anglo-centric West, the discipline faces a severe bias problem.

The remainder of this chapter discusses the objects of study, methods and research questions of sociolinguistics with a view to the changes brought about by the digital turn in general, and the institutional formation of the digital humanities in particular

Sociolinguistics and the digital turn

Objects of study

Sociolinguists' insistence that theory formulation be based on observable language use has inevitably inspired discussions around the definition, preservation, management and interpretation of data. Even before the digital turn, issues typically associated with the digital humanities received detailed attention in sociolinguistics. For instance, one of the first machine-readable language corpora ever compiled, the Montreal Corpus (Sankoff and Cedergren 1972), was collected with the explicit goal of studying sociolinguistic variation. This project pioneered not only technical aspects such as data storage and access but also ethical discussions that have become increasingly prevalent in times of social media and big data (see Carter et al. 2016 for a study of emerging attitudes and practices among academics and in institutions with respect to the ethical issues pertaining to the use of online data). Other corpora, such as the foundational Brown Corpus (Kučera and Francis 1967) and its many intellectual successors,² although not primarily compiled for sociolinguistic purposes, have seen wide application in the study of language variation as well (e.g. Andersen and Bech 2013).

Developments in communications technology, an explosion in computing power and simultaneous decline in hardware costs and increasing data storage capacities have brought drastic changes to the sociolinguistic 'datascape' of the twenty-first century. At the most trivial level, these are developments of scale and access: digitised language data of increasing magnitude is becoming more easily available for more people, often at the click of a mouse. Where earlier corpus compilers faced the task of manually collecting, transcribing and annotating their material, most of these steps are (semi-)automated in today's megacorpora projects. The Corpus of Global Web-based English (GloWbE, Davies and Fuchs 2015) is a case in point: a search algorithm harvests web pages from the English-speaking Internet, categorises them according to their country of origin and automatically annotates each word with part-of-speech information. The result is a corpus roughly 20,000 times the size of Brown, which can be searched through a convenient web interface. Such an increasing automatisisation is not without its problems, however. We do not discuss these with regard to GloWbE in particular (but see e.g. Mair 2015; Nelson 2015), but return to the general issue in the section on research questions.

Apart from this explosion in corpus size, advances in digital technology have also enabled the creation and dissemination of new kinds of corpus data. Whereas formerly, anything beyond the level of machine-readable text would have posed serious problems in terms of storage and data handling, speech and video archives are being built at an increasing rate. This concerns both legacy data – archival material collected at an earlier stage that is digitised and made more widely available in this form – and newly curated speech and video archives for sociolinguistic analysis. The International Corpus of English (ICE) project (Greenbaum and Nelson 1996), for instance, began as a purely text-based corpus (featuring spoken language in orthographic transcription only). Now, however, several of its national sub-corpora provide access to the full audio recordings for their spoken components. In the case of ICE Nigeria (Wunder et al. 2010), the recordings are even fully time-aligned to the corresponding text files. The possibility of hosting

digital speech archives is particularly relevant for the documentation of lesser-known, uncodified or endangered languages and varieties.

Researchers are largely in agreement that sociolinguistics stands to gain much from the availability of increasingly large datasets and sophisticated computational methods of analysis (Szmrecsanyi 2017), and particularly from ‘sociolinguistically rich, unconventional corpora’ (Kendall 2011: 383) that reflect especially informal spoken language in greater variety than previously. As the editors of this handbook pointed out to us, there would also be potential drawbacks: for example, if large corpora were to cause sociolinguists to take computer-assisted armchair views of language variation and to abandon personal interaction with speech community members. In this way, an important source of insight into both the social functions and the socially conditioned variability of language would be lost. To counterbalance this dystopian vision, we raise two points:

- 1 Along with the changes affecting sociolinguistics in recent years that have to do with computational infrastructure and methodology, the most significant change in the field has been an orientation towards more speaker-based, interactionally and community-oriented study designs. Often, both trends are combined.
- 2 The kind of work that, in this dystopian view of the link between corpora and sociolinguistics, is seen as replacing fieldwork-based sociolinguistics is not actually sociolinguistics proper. In Szmrecsanyi’s presentation,³ it is instead ‘corpus based variationist linguistics’, a sub-type of corpus linguistics which lacks precisely the interest in interactional context and observation that makes sociolinguistics sociolinguistics (see Figure 16.1). A *replacement* of sociolinguistics by corpus-based variationist linguistics is therefore not just unlikely but also logically impossible.

At least as significant as the recent explosion in easily harvestable language data are new kinds of data resulting from new contexts in which language use is implicated. Communication in a wide range of technological modes, such as weblogs, email, Facebook, Twitter, etc., has been summarised under the umbrella term digital discourse (Thurlow and Mroczek 2011). While lay commentary, but also some early scholarship, favours a totalising and exoticising perspective on such instances of language use – presenting them as internally undifferentiated, maximally distant from offline language and often structurally deficient (Thurlow 2006; Squires 2010) – sociolinguistic research has documented its regularity (Tagliamonte and Denis 2008), communicative adequacy (Jones and Schieffelin 2009) and the complex interplay of technological constraints and affordances with communicative needs and linguistic forms in digital environments (Herring and Androutsopoulos 2015).

Yet many sociolinguists’ interest in digital discourse has been ambivalent at best. On the one hand, there is great promise in the availability of terabytes of user-generated language data. On the other, research on language variation and change has traditionally emphasised the vernacular as its privileged source of data, defined as the style ‘in which the minimum attention is paid to speech’ (Labov 1984: 29). This emphasis does not always sit comfortably with the highly planned and rhetorical character of some instances of digital discourse (Hinrichs 2018). A loosening of time constraints during production, an awareness of multiple potential audiences (Marwick and boyd 2011) and the competition for attention in a noisy environment result in sometimes highly strategic language use whose relationship to a vernacular baseline is not always straightforward.

Consequently, rather than drawing on digital discourse to make statements about language in general, researchers have focused on the specific situational, technological and

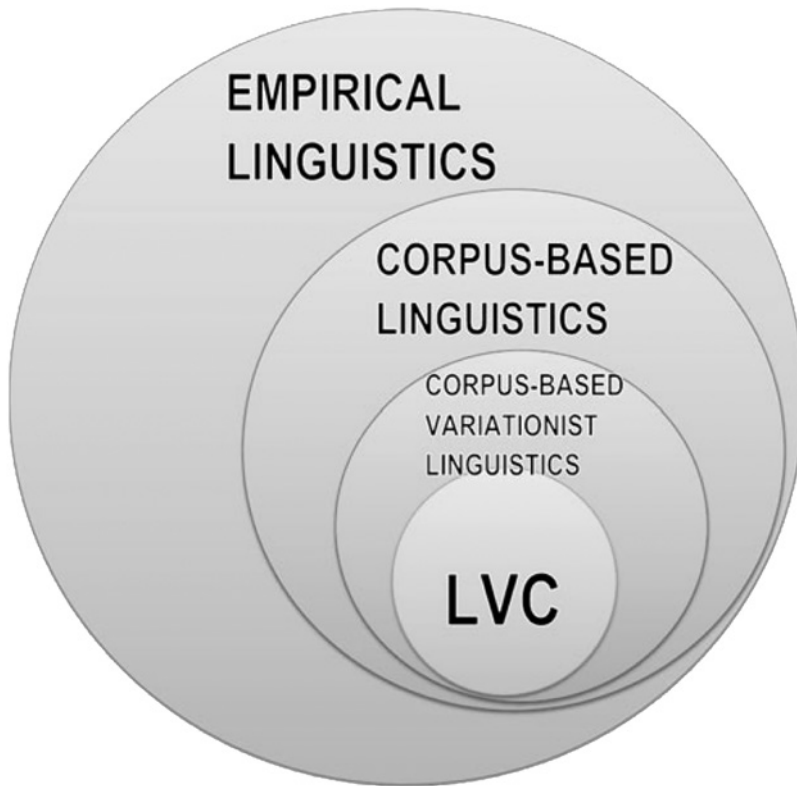


Figure 16.1 Sociolinguistics (here represented as 'LVC', i.e. 'language variation and change studies') versus other branches of empirical linguistics

Source: Szmrecsanyi 2017: 687

structural features of individual communicative platforms. Among the most often-discussed linguistic characteristics of digital data are innovative orthographic conventions such as strategic lengthening (as in *loooooove*), creative use of capitalisation, the replacement of phonetic sequences with homophonous numbers (Blommaert 2011) or use of emoticons and emojis (Schnoebelen 2012). The diminished normative influence of the orthographic standard is an aspect of many forms of computer-mediated discourse (CMD) that makes for interesting new research questions, but also poses problems to established techniques in natural language processing, such as automated part-of-speech tagging (Eisenstein 2013). In addition, it is now common to encounter languages and language varieties in writing that before the digital turn had been largely restricted to oral communication within their respective communities, such as Jamaican Creole (Hinrichs 2006) or Nigerian Pidgin (Heyd and Mair 2014). What all of these developments contribute to is a new level of informality in writing.

A further important aspect of digital language data are the high amounts and often new forms of multilingualism that can be found on the World Wide Web and in social networking sites. Instant communication across the globe can bring geographically disconnected languages (see Danet and Herring 2007), but also different varieties of English into contact with each other (Mair 2013). Such processes do not only produce new kinds of multilingual

data, but these data trouble any straightforward link between place, community and language varieties. In such cases, place as well as language are best understood as socially constructed resources that speakers draw on in identity performances rather than merely reflections of their demographic and regional position (Heyd and Honkanen 2015).

Outside of the arena of the (type-)written word, linguistic data are produced for communication along audio, visual and audio-visual channels in which new constellations of participants and technological affordances spawn novel linguistic and communicative behaviours. Cases in point are Skype conversations (Licoppe and Morel 2012), multiplayer online games (Iorio 2012; Keating and Sunakawa 2010) or videos made specifically for YouTube audiences (Kiesling 2018; Chun 2013). In these kinds of data, language is embedded into multimodal contexts that require attention to the combination of different communicative channels and semiotic modes, and participant constellations go beyond the typical settings of face-to-face, offline communication. Often as well, multimodal language data can be downloaded and electronically manipulated by users or embedded into new contexts, such as when users remix or dub over video material from YouTube (Bohmann 2016; Androutsopoulos 2010). Processes of de- and re-contextualisation and remediation therefore are important aspects of digital sociolinguistic data (Heyd 2015; Squires 2014).

As the previous discussion shows, the concept of data is an established and central one within sociolinguistics. By comparison, the humanities at large have only fairly recently incorporated the term into their conceptual arsenal, and this orientation towards data is directly related to the rise of the digital in the humanities. The central role that data play in present discussions is underscored by publications such as the *Journal of Open Humanities Data*. This rise in interest is accompanied by a discussion about what counts as data in the humanities. According to Borgman (2015), anything a researcher presents as evidence in support of an argument fits the bill. However, many scholars oppose such an all-inclusive perspective. Drucker (2011) argues that the term itself is problematic on etymological grounds, since Latin *datum* means ‘that which is given,’ and opts for a new terminology entirely. Schöch (2013) emphasises the abstraction inherent in data, opposed to the context-richness and situated-ness of traditional humanities arguments, and defines data as ‘a digital, selectively constructed, machine-actionable abstraction representing some aspects of a given object of humanistic inquiry’. At the heart of the discussion, then, is a tension between the attraction of *big* data and the value of close, contextually informed analysis typical of humanities research, whether this opposition is understood as one of degree (*pace* Borgman) or of kind (*pace* Schöch).

Similar debates have been going on in sociolinguistics since its earliest days and have risen to prominence in the context of the ‘third wave of variation studies’ (Eckert 2012), which shifts the focus of attention from large-scale demographic patterns to situated stylistic variation. Sociolinguists occupy different positions along the spectrum between data abstraction, pattern-based research and detailed, contextualised, ethnographic analysis, and there is certainly contention among these perspectives. In general, however, the discussion is not engaged in absolute terms. Most scholars recognise that there are drawbacks and advantages to both quantitative and qualitative approaches, and see them as mutually informing each other, rather than in strict opposition. Some of the more influential sociolinguistic work in recent years has arguably developed from a flexible, methodologically varied understanding of how to approach sociolinguistic data. For example, Mendoza-Denton (2008) studies a community of high school gang members from an ethnographic point of view. The data she considers for her study include traditional interviews between the researcher and participants, but also drawings, handwritten notes, photos, film, etc.

Methods

The diversity of perspectives mentioned earlier also pertains at the level of methods in sociolinguistics. Borrowing from a range of neighbouring disciplines, and taking different views on the nexus between the linguistic and the social, sociolinguistics typically takes pride in its methodological pluralism. Consequently, there are multiple intersections between sociolinguistics and ‘the digital’ at the level of methods. These can be grouped into three types.

The first type consists of the application of traditional sociolinguistic methods to new, digital data of the kind outlined earlier. Conversation analysts, for instance, have been studying the sequential organisation of turns-at-talk since the late 1960s (Sacks et al. 1974). However, many of the features of real-time, face-to-face conversational organisation receive a different inflection in digital environments with new participant constellations, new dimensions of (a)synchronicity and new semiotic means for the organisation of interaction. Established conversation analytical methods have played an important role in understanding the systematicity of many forms of digital communication and in relating this systematicity to that found in offline communication (Rintel et al. 2001; Stommel 2008; Spilioti 2011). Similar observations can be made for the structure of participation frameworks and production formats. Goffman’s (1974, 1981) work in this regard has found productive application in the context of remediation and remixing on YouTube (Bohmann 2016), socialisation into online communities (Weber 2011) and participant alignment in private emails (Georgakopoulou 2011).

At the most general level, digitally mediated communication is discourse, that is, ‘meaningful semiotic human activity seen in connection with social, cultural, and historical patterns and developments of use’ (Blommaert 2005: 3). As such, sociolinguistic methods of discourse analysis have been adapted and applied in computer-mediated contexts. Herring (2004) is a foundational text in this regard, elaborated and updated to account for more recent technological developments in Herring and Androutsopoulos (2015). CMD analysis has developed into a prolific and variegated sub-field in its own right, with publications such as the online journal *Language@Internet* and edited volumes such as Thurlow and Mroczek (2011), Georgakopoulou and Spilioti (2015) and, for English specifically, Squires (2016).

The second type of intersection between digital technology and sociolinguistic methods consists in recruiting (more elaborate) digital technologies to engage traditional sociolinguistic data. This primarily concerns quantitative statistical analysis, but also automatic alignment, annotation and categorisation of data. Compared to many other disciplines in the humanities, sociolinguists have started looking to digital technology for methodological support early on. The variationist focus on the relationship between demographic categories and aggregate patterns of linguistic variation effected the introduction of statistical procedures into the research paradigm. As early as 1974, VARBRUL (Cedergren and Sankoff 1974) was developed as a computational analysis tool for sociolinguistic data. Similarly, corpus linguists of the first hour developed computational methods for searching, concordancing and finding collocations of words in a corpus.

As technology developed, so did the digital sociolinguistic toolkit. In terms of predictive statistical analysis, the VARBRUL programme went through several iterations, alterations and modifications to remedy some of its initial shortcomings and adapt to newly available technology. While the basic method at its heart – logistic regression – has remained the same, there has been a constant improvement in terms of ease and flexibility of its application. The most notable development in recent years has been the incorporation of mixed-effects (linear and logistic) regression modelling provided by the Rbrul software (Johnson

2009). At the same time, open-source statistical analysis software like R (within which Rbrul is implemented) nowadays is available to any researcher with a personal computer and the required coding skills, making statistical analysis more flexible and customisable than any out-of-the-box programme could provide. Consequently, many sociolinguists of the present generation work directly in R. Regression modelling remains the most important statistical method, but is now being supplemented by other approaches, such as random forests and conditional inference trees (Tagliamonte and Baayen 2012) or generalised additive models (Tamminga et al. 2016). In addition to these methods of analysis, statistical software facilitates graphical data exploration and visualisation.

In corpus studies of language variation, similar tools are becoming widespread. Like variationist sociolinguists, many corpus researchers now work directly in R. Among the frequent statistical methods found in this research paradigm, in addition to the ones outlined earlier, is the calculation of key words (Archer 2009), that is, those words that are particularly (in)frequent in a corpus compared to a reference corpus. Another method, which in fact was first applied in the 1980s (e.g. Biber 1988), is the use of factor analysis and similar clustering techniques (Moisl 2015). In general, such procedures allow the description of multidimensional objects in terms of a smaller set of abstract dimensions and can be used to quantify similarities and differences between texts, corpora or varieties. The case study we present below illustrates this procedure in more detail.

Technological developments have also significantly affected the treatment of spoken language data. Digital recordings can be visually inspected and analysed with the software PRAAT (Boersma 2001), which also supports automatisisation of many analytical steps by means of scripts. Transcription programmes such as ELAN (Brugman and Russell 2004) help align transcription glosses with time-stamped audio files. More recently, advances have been made towards fully automating the process of phonetic transcription and time alignment at the level of individual sounds with FAVE (Rosenfelder et al. 2011) and DARLA (Reddy and Stanford 2015). While the latter two services are based on dictionaries of Standard English phonemic word transcriptions, promising advances are being made in the area of dictionaryless time alignment, which can be used with non-English speech recordings and their orthographic transcripts, cf. the WebMAUS tool (Kisler et al. 2017), available freely online.

Finally, there are tendencies towards a more direct conversation between computational linguistics and sociolinguistics, reflected in the coining of the new term ‘computational sociolinguistics’ (Nguyen et al. 2016). In almost all cases, research of this kind is based on large-scale social media data and consequently belongs to the third type of intersection discussed here: a sociolinguistically informed application of state-of-the-art, digital methods to new kinds of digital language data. A number of studies use computational procedures that are normally employed in the analysis of offline speech data in order to understand patterns of variation in online linguistic behaviour, thus leveraging the availability of such digital data, as well as explicitly connecting their findings to known tendencies from earlier research (e.g. Eisenstein 2015; Iorio 2012; Tagliamonte and Denis 2008). Similarly, large samples of online language are helpful in triangulating findings from more controlled, small-scale studies (e.g. Haddican and Johnson 2012).

Beyond this expansion of quantitative scope, the size and nature of digital data – often from social media communication – also allow for the incorporation of new methods from computational linguistics. A large range of machine learning techniques⁴ have recently been enlisted in the social analysis of language, such as support vector machines (Fink et al. 2012), naïve Bayes (e.g. Goswami et al. 2009) or topic modelling (Schnoebelen 2012).

Research from a computational linguistic perspective tends to take a static view on social dimensions as fixed inherent properties and has typically pursued the goal of predicting age, gender and other categories from language use. Sociolinguistics, on the other hand, regards social categories as at least partly fluid and is interested in how the social structures vary in language. However, some recent work has successfully surmounted this theoretical divide. Bamman et al. (2014) use a range of machine learning techniques to problematise any direct link between a Twitter user's gender and their linguistic performance, highlighting instead the importance of stylistic agency and friendship network structure (see also Schnoebelen 2012 for work in a similar direction).

Some of the new computational methods in sociolinguistics are particularly driven by the kind of data available for analysis. For instance, the explicit coding of network information in social media data has led to new statistical methods for the analysis of linguistic accommodation (Jones et al. 2014) and audience design (Pavalanathan and Eisenstein 2015). Another aspect of social variation that is witnessing renewed methodological interest is geography. Precise geo-information in many kinds of social media data, in combination with their sheer quantity, facilitates new ways of studying language variation across space (e.g. Huang et al. 2016; Grieve et al. 2013). These developments are contributing to the field of dialectometry by combining quantitative-computational methods with the research interests of traditional dialectology (see Nerbonne and Kretzschmar 2003; and for English in particular Grieve 2016; Szmrecsanyi 2013).

As this discussion has shown, digital technology has an important role to play at the level of method in sociolinguistics, and has done so since the earliest days of the discipline. Current advances in digital technology are highly relevant to the field, both in facilitating and increasing the scale of established methods and in bringing new tools to the table. However, unlike elsewhere in the humanities, the role of digital methods is not perceived as bringing about a paradigm shift as much as contributing to a constant development of the research agenda.

Research questions

It is difficult nowadays to find any sociolinguistic work that is not influenced by digital technology in one way or another. For many researchers, especially those of the variationist school, spontaneous spoken interaction in community settings remains the most important object of study. But even these scholars engage their traditional research questions with increasingly sophisticated digital methods at every step of the analytical procedure, from data collection to annotation, feature extraction, and on to statistical analysis.

However, new research questions have also emerged on the basis of new kinds of digital language data. An important focus in this regard is the relationship between variation found in these new environments and that which has been studied at the spoken community level. Against widespread, but unfounded, reports of a supposed decline in linguistic standards brought about by digital communication, Tagliamonte and Denis (2008) demonstrate that variation in instant messaging among teenagers follows the same patterns of structured heterogeneity it does in their speech. On a global level, Davies and Fuchs (2015) are able to replicate many findings from small-scale analysis of carefully controlled offline language use in their analysis of GloWbE.

Despite its regularity, computer-mediated communication tends to trouble a clear distinction between spoken and written language as separate, if related, systems. Earlier sociolinguistic research was concerned with the question of how 'oral' or 'literal' (a certain type of)

CMD is (e.g. Yates 1996). While this simplistic question has given way to a more dynamic, nuanced understanding of variation in particular ‘socio-technical modes’ (Herring 2002), more specific aspects of the relationship between offline speech and online writing remain highly relevant. For instance, the prevalence of non-standard written language on Twitter has made it possible to study ‘phonologically-motivated orthographic variation’ (Eisenstein 2015; see also Callier 2016), where variation in the phonetic shape of a word is rendered through spelling choice, such as dropping the *g* at the end of *-ing* sequences. Beyond individual variables, entire varieties without a widely accepted codified orthography are now being represented in written form by computer users. Investigating the patterns of variation found in these emerging grassroots orthographies brings up a number of research questions having to do with the understandings participants have of their vernacular and the way they relate it to Standard English (Jones 2016; Hinrichs and White-Sustaíta 2011).

Other orthographic choices are not directly motivated by variation in spoken language, but nonetheless constitute elements of linguistic agency that have received detailed sociolinguistic attention. These concern strategies such as emoticon use (Schnoebelen 2012; Vandergriff 2014) or the substitution of certain character strings by numeric homophones (Blommaert 2012). The extent to which such choices are systematic and regular, the interactional purposes they are enlisted for and differences in different groups’ usage are among the new research questions resulting from this kind of variation. Thus, digital data contribute significantly to an emerging sociolinguistics of writing (Blommaert 2013). A related question, arising from the search- and citability of social media language (Zappavigna 2015), is that of its uptake and embedding in the wider media and text type ecology (Squires and Iorio 2014).

The ‘mediated’ aspect of CMD as well is a productive source of new research questions. Any form of digital communication is subject to constraints and affordances of the specific communicative infrastructure it is embedded in, both in terms of technology and social conventions. An important strand of digital sociolinguistic research investigates the relationship between medium and observable forms of communication. Among the specific media that have received attention in this context are Internet Relay Chat (Rintel et al. 2001), text messages (Spilioti 2011), weblogs (Herring and Paolillo 2006), Twitter (Callier 2016; Jones 2016; Eisenstein 2015) and Facebook (Akkaya 2012). Research questions in this area revolve around the extent and nature of the ways that design elements determine linguistic patterns, but also the ways in which users appropriate technology and develop new conventions and expressive strategies. A case in point is synchronicity, that is, the extent to which communication takes place under conditions of real-time co-presence of sender and receiver (Herring 2007). This is largely considered a technologically determined feature of the communicative situation, but under certain circumstances, participants may develop strategies to reframe a technologically asynchronous medium as a synchronous one (Bieswanger 2016).

Another focus for new research questions is the relationship between linguistic variation and space, whether understood as physical geography or the territory associated with a linguistic community. Digital communications technology opens up new possibilities for interactions across distances on the global scale, bringing previously unconnected people and linguistic resources into contact with each other. The result is both a loosening of the link between language and place and often also the rise of hybrid forms of language, drawing on resources from several origins (Heyd and Mair 2014). Such processes have led some sociolinguists to propose a ‘post-varieties approach’ (Sergeant and Tagg 2011), which rejects the theoretical fiction of languages as bounded entities.

While many in the research community acknowledge the fictional nature of concepts such as language or variety, there is current debate over whether this fiction is a convenient and productive one or whether it should be discarded entirely. Irrespective of which stance scholars take on this issue, it is hard to deny that digital communication spawns certain kinds of linguistic behaviour that are difficult to account for with face-to-face, speech community models of variation. Numerous studies analyse the formation of new online communities of practice, often more fleeting and more heterogeneous than traditional speech communities. In these contexts, language becomes a central resource in identity construction (Peuronen 2011; Stommel 2008), and the social construction of place becomes an important project (Kiesling 2018; Heyd and Honkanen 2015). The role of vernacular forms in the creative performance of online identities invites attention to ‘vernacular spectacles’ (Androutsopoulos 2010).

Fewer entirely new sociolinguistic research questions have developed out of the availability of new, digital methods. One relevant innovation is the fact that larger amounts of data can be collected to cover a wider geographical area than previously feasible. Consequently, while most traditional sociolinguistic studies centred on individual, geographically defined communities, making comparisons across space is now a more realistic project.

At the highest theoretical level, sociolinguistic research is in a position to critically interrogate *the digital* with an understanding of social theory and the central role of language in both. In this sense, sociolinguists can contribute to critical data studies. The ability to store, process and analyse very large amounts of data – and language data are crucial among these – increasingly represents the key regulating criterion in economic as well as political competition among different actors. The fascination of big data and the promise of ‘better tools’ have at times directed attention away from the inequalities arising out of the digital turn and the pressing ethical and epistemic questions that arise from it.

At the level of research and knowledge creation, boyd and Crawford (2012) point out that access to as well as the expertise required to analyse big data are highly unequally distributed, resulting in a clear divide between ‘data haves and data have nots’. The former are typically associated with corporate or government interests, while universities are rarely able to provide data resources at competitive levels. This means that the purposes big data are utilised for are highly specific and the possibility for external control is limited.

At a more theoretical level, a critical perspective is warranted on the kind of solutions offered by big data in the first place. The scholarly discussion at present is sometimes carried by the fantasy that control over information will lead to a privileged position of knowledge and control in general (Posner 2012). However, this appears to be an illusion on several counts. First, in algorithmic systems of rising complexity, it becomes increasingly difficult for any individual agent to have a full understanding of the system and the information contained in it. But the system itself aspires to full control, and consequently power and agency is wrested from human agents and resides in the algorithm itself. Second, while data are getting bigger and bigger, the quality remains up for debate. Irrespective of what specific shape data take, they constitute an abstraction away from human experience and the contextual niceties of specific social situations. Both these points suggest the possibility that data are fundamentally dehumanising. It remains a central task of the digital humanities, and sociolinguistics along with it, to establish human perspectives on big data. The focus on social theory and on interpretation that separates sociolinguistics from computational linguistics (Nguyen et al. 2016) is not a hurdle to overcome, but a necessary corrective that may safeguard us from stepping into a new algorithmic no man’s land. Similarly, while mega-corpus projects such as GloWbE (see earlier) are generally well-received, there is a healthy level

of scepticism as to their representativeness and balance (Mair 2015; Nelson 2015). Many researchers today incorporate ‘big and messy’ corpora into their analyses, but triangulate and balance their strengths and weaknesses with those of ‘small and tidy’ ones (Mair 2006).

Not included in the earlier discussion is the perspective of the users themselves who generate the bulk of big data. Legal frameworks for access to, and distribution of, digital (primarily social media) data are not yet fully developed, and it is important to recognise ethical issues beyond staying on the safe side of the law (boyd and Crawford 2012). While tweets, for instance, are public in legal terms, there is clear evidence that users often have only a hazy understanding of who may or may not read their messages (Marwick and boyd 2011). As soon as data are stored and shared, unforeseeable ethical questions in their future use need to be considered. Corpus linguists were among the earliest humanities researchers to engage with such questions, and issues of privacy remain central to many corpus-building enterprises.

Finally, how digital communication infrastructures regulate and influence communicative patterns is a question in need of serious attention. The tendency for many platforms to prioritise content a user already agrees with has been criticised as creating echo chambers or ‘filter bubbles’ (Pariser 2011) that isolate users from opinions diverging from their own. While the focus in sociolinguistics has traditionally been strictly on the structural aspect of language use, not on content, the tradition of social network analysis in the discipline (Milroy 1980) has the potential to empirically address issues of political polarisation in social media and similar echo chamber effects. Questions of this kind are not currently at the top of the sociolinguistic agenda, and more engagement with them is needed.

A case study of digital data, using digital methods: Relations among text types and varieties in twenty-first-century English

In what follows, we illustrate the points we made earlier with a case study. The research presented here is part of Bohmann’s (2019) book-length study. It uses both digitised and born-digital data and employs computational methods in order to address sociolinguistic questions that are particularly relevant in the globally connected world of computer-mediated communication. Within sociolinguistics, this research is situated in the context of world Englishes research and combines a corpus linguistic with a lectometric perspective. The motivation for the project is to develop a multidimensional system of linguistic variation (cf. Biber 1988) that can account for and quantify differences among national varieties of English as well as among different communicative genres. As space limitations do not permit a summary of the entire research, we restrict the discussion here to one dimension of variation the project establishes.

The project departs from the recognition that language varieties are highly multidimensional objects: they display variation along a very large number of variables on different levels of linguistic description. On the one hand, this means that any accountable study should attempt to measure as large a portion of that variation as possible. On the other, it is clearly unproductive to consider each variable in isolation. The goal, instead, is to develop robust general tendencies out of many individual variables.

A solution to the first point requires fairly large amounts of data and the possibility of automatically extracting frequency information on many features simultaneously. Here, both traditional corpora and samples from social networking sites have an important role to play. This case study specifically draws on data from ten national components of the International Corpus of English (ICE; Greenbaum and Nelson 1996). For each nation, written

and spoken language samples from a range of genres – from spontaneous conversations to scripted speeches to academic writing – are included. The total size of the corpus data is 1 million words per variety, totalling 10 million words.

This is supplemented with geolocated language data from Twitter,⁵ sampled from across the English-speaking world. While it is possible to collect such data at much greater orders of magnitude than what is represented in ICE, the goal of maintaining a genre balance warrants self-imposed restrictions on the number of tweets included in the study. In total, the Twitter corpus is composed of just over 6.8 million words. These are bundled into multi-tweet texts categorised by country code and time of writing in order to ensure comparison with the 2,000-word samples of ICE.

Table 16.1 gives an overview of the size and genre composition of the data. The data are evenly distributed across the ten national varieties, with two exceptions. First, ICE-US only contains the written part of the ICE sampling framework; therefore, no spoken data for American English are included in the present study. Second, it was difficult to sample appropriate amounts of Twitter data for smaller countries or those where Twitter is less extensively used. The total number of text samples for the ten varieties therefore range from 220 (USA, GB) to 53 (HK). Additional smaller varieties have been included in the Twitter corpus to make up for this skew and to represent the full breadth of global variation in English Twitter.

For each individual text sample in the combined ICE and Twitter corpus, frequency information about each of 236 linguistic variables is extracted. A complete description of the entire feature catalogue is beyond the scope of this chapter, but they are predominantly morpho-syntactic (e.g. frequency of the passive voice), discourse-organisational (e.g. frequency of adverbial subordinators) and, to a lesser extent, lexical (e.g. frequency of discourse-structuring *so*) and orthographic variables (e.g. non-standard spelling of definite articles). Extraction proceeds automatically by means of a Python script that relies heavily on regular expression searches.⁶

Next, it is necessary to account for all the measured variation in terms that reduce its complexity to human-processable dimensions without sacrificing too much information. For that purpose, we use the statistical technique of factor analysis (Thompson 2004). Factor

Table 16.1 The combined ICE and Twitter corpus in Bohmann (2019)

<i>Genre (Code)</i>	<i>N/texts</i>	<i>N/words</i>
private dialogue (S1A)	900	1,962,475
public dialogue (S1B)	720	1,594,650
spontaneous monologue (S2A)	630	1,396,554
scripted monologue (S2B)	450	979,624
student writing (W1A)	200	425,028
correspondence (W1B)	300	639,255
academic prose (W2A)	400	885,417
informational writing (W2B)	400	872,954
news reports (W2C)	200	426,526
instructional writing (W2D)	200	433,965
persuasive writing (W2E)	100	212,738
novels and stories (W2F)	200	441,125
Twitter messages (TwICE)	2,609	6,855,727
Total	7,309	17,126,038

analysis is a latent variable technique that describes covariance found in measured variables in terms of a smaller set of more general variables (also called factors or dimensions). The output of a factor analysis is simply a pre-selected number of factors, along with measurements of association (structure coefficients) for each combination of a factor and a measured variable. The researcher uses a pre-defined cut-off criterion for determining the measured variables that relevantly contribute to a factor, in our case all variables with a structure coefficient above 0.3 or below -0.3 . For reasons of constraints of space, we do not discuss the many technical decisions involved in choosing an appropriate factor solution here, nor do we present the solution in its entirety (but see Bohmann 2019). The model consists of ten dimensions of variation, each of which comprises between 5 and 64 measured variables. Here, we demonstrate the analytical steps that lead to the interpretation and labelling of one dimension.

The two most important measured variables on this dimension (factor 5 in Bohmann 2019), with structure coefficients of 0.61 and 0.60 respectively, are preposition sequences and place adverbials, such as *here*, *nearby*, *away*, etc. Next in salience are time adverbials (*again*, *yesterday*, *early*, etc.), with a structure coefficient of 0.42. All of these features represent strategies to demarcate physical relations, either in the spatial or the temporal domain. This is evident in the case of the two adverbial classes and can easily be shown for preposition sequences as well. These comprise members like *from within*, *out onto*, *up against*, etc., that typically put two noun phrases into spatial (and sometimes temporal, as in *over with*) relation.

The remaining features, all with smaller structure coefficients, are *gh*-sequences (0.35), mean word length (-0.32) and nominal suffix *-ity* (-0.31). These have no direct connection to the expression of physical relations. Rather, they all serve as indicators of a style that avoids structural complexity: by choosing shorter words on average, by preferring Germanic words (*gh*) over Latinate ones (*-ity*) and by disfavours nominal expression of ideas (*-ity*). Taking the primary dimension (expressed by the adverbial classes and preposition sequences) and the secondary one (structural simplicity) together, it appears that high factor 5 scores are indicative of texts with a low focus on tight informational integration that explicitly code physical situational properties.

To further explore and corroborate this intuition, it is possible to calculate a factor 5 score for each individual text in the corpus. Factor scores can then be compared for individual text samples as well as for different genres. Figure 16.2 gives an indication of how factor 5 stratifies the texts in the corpus. Most noteworthy is the lack of a clear distinction between spoken, written and Twitter modalities. Most genres' mean factor 5 scores cluster around values of 0, and the two highest-scoring genres represent different modalities: unplanned monologues and fiction writing, with mean scores of 0.65 and 0.67, respectively.

A closer look at the spontaneous monologue category reveals further patterns. While this category on the whole is not radically different from the others, there is a sub-group within it that is highly distinct, indicated by the high numbers reached by the upper whisker of the boxplot for this genre. Indeed, an ordinal ranking of individual corpus texts along factor 5 reveals that the highest-scoring 69 texts are all from one specific subcategory and that only 5 out of the 111 highest-scoring ones do not belong to this category: live radio commentary on sports events. This kind of communicative genre has some highly specific properties that help account for its separate position along factor 5. Speaker and audience are physically separated, but communication takes place in real time. The flow of communication is dictated by the succession of events to which the audience has no access, other than by means of the commentator's discourse. This accounts for the predominance of physical anchoring

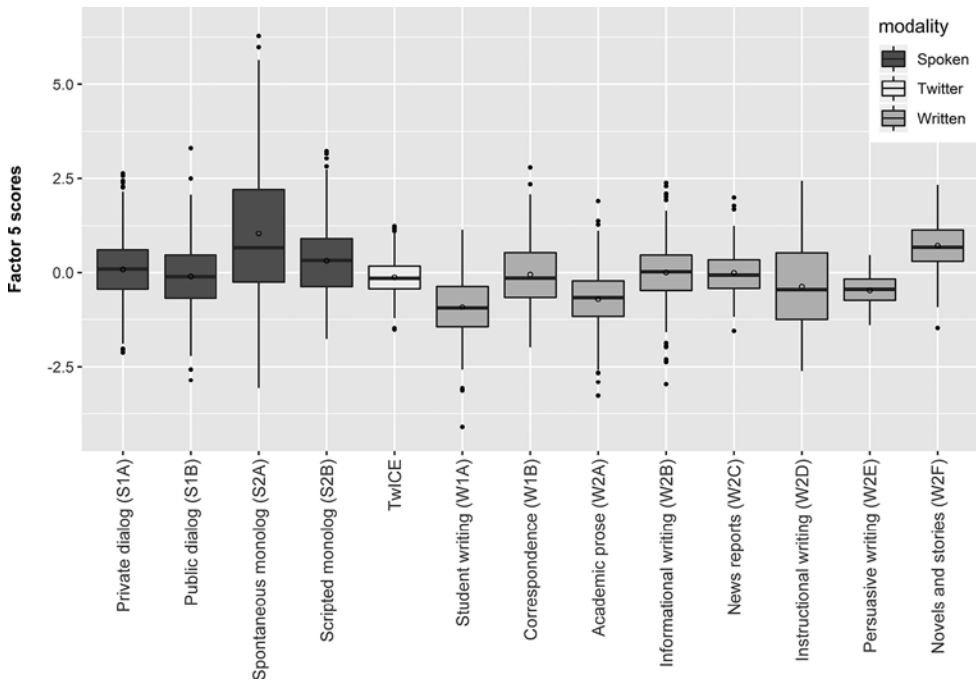


Figure 16.2 Factor 5 score distribution for different genres

strategies such as time and space adverbials. To various degrees, sports events proceed at a relatively fast pace, requiring simply structured language to allow the commentator simply to keep up with the temporal progression of the event. This accounts for the negative coefficients of mean word length and *-ity* suffixes, and potentially also for the positive one in the case of *gh*-sequences.

Having established the basic dimension of variation covered by factor 5, it is instructive to consider a text that is not sports reporting but incorporates relevant features to a high degree. Example (1) is taken from a fiction text:

- (1) Inside the house now and so many people. They all seem to know her, but she only knows some. Jack comes up the brown stairs, a tray over his head, shouting 'Load coming through' again and again. And all the old Aunts laugh and say he's a scream. There 's sherry and port wine for each of the ladies and stout for the men and whiskey as well. And each glass grows its own deep colour from stem up to brim. <ICE-IRE-W2F-019>⁷

This passage features one salient factor 5 variable (underlined) per 9 words, bringing it into close stylistic proximity to sports commentary. Like the latter, (1) is characterised by a fairly fast progression of events, involving various actors (the protagonist, Jack, the aunts and unspecified number of ladies and men, as well as inanimate objects like glasses) moving through or being situated in space. The effect is one of immediacy and potentially a disorientating number of things happening simultaneously.

This case study exemplifies several aspects of sociolinguistics in the digital age. In terms of data, it depends both on digital post-processing and storage of traditional linguistic material (in the form of ICE) as well as on access to new kinds of language produced in the digital environment of Twitter. And while size and availability of data are important aspects in sociolinguistic research, the self-imposed size limitations on the Twitter component of the corpus demonstrate that bigger is not necessarily always better. Theoretical aspects such as balance and representativeness need to be kept in mind. The choice to sample tweets in this study was not primarily motivated by access to masses of data, but was done in order to represent an increasingly prevalent shade of the present-day media ecology that is not included in the ICE sampling framework.

On the methodological plane, the study draws on a range of different digital techniques, from harvesting tweets from Twitter's application programming interface (API) to using Python scripts for feature extraction and R for statistical analysis and data visualisation. All of these methods depend on technology that is currently available on personal computers, but not too long ago were beyond the scope of even ambitious computer centres. The analytical progression from large-scale aggregate patterns to more contextualised, situated examples of language use may also be taken as exemplary for a field that has strong traditions in quantitative and qualitative analysis.

This case study also makes more tangible some of the ways in which sociolinguistically informed research may contribute to related fields in the digital humanities. The finding that certain samples of narrative writing assume a stylistic profile similar to live sports reporting may be of interest to literary scholars and the method demonstrated here could well support distance-reading approaches that automatically identify such passages in an author's work, or make differences between individual authors or works measurable. Students of political or institutional discourse, of pedagogy and other fields may likewise be interested in similar ways of quantifying certain aspects of the textual data they study. The sociolinguistic toolkit offers solutions in these cases.

Future directions

Sociolinguistics is fully engaged with digital data, methods and research questions in many ways already. Still, there are new intersections that remain to be explored and refined more fully and continuing technological developments that will drive the development of the field in the future.

One important new trend is the large-scale analysis of linguistic material in modalities other than the written word. Corpus projects and analytic tools, such as those mentioned earlier, are gradually becoming available, and some initial important research on their basis exists already. However, most analyses of spoken data remain small-scale, survey-based community studies, while large-scale studies almost exclusively rely on written data.

Another conversation that will likely gather momentum in the near future is that between computational linguistics and sociolinguistics. The sophisticated statistical and computational tools of the former may benefit the latter to a great extent and, conversely, a better understanding of the social dynamics of language use can serve to improve machine learning tasks. A more effective synergy is still hampered by the different theoretical orientations of the two fields. Computational linguists take a sceptical perspective regarding theory and focus their aims on task-specific performance and predictive accuracy. Sociolinguists, on

the other hand, focus on grounding empirical findings in social theory and ask to what extent patterns in the data are conceptually valid and interpretable (Nguyen et al. 2016). There is also tension between the social constructivist understanding many sociolinguists have of identity categories like gender or ethnicity and the essentialist view typically adopted in computational linguistics. Studies like Bamman et al. (2014) or Schnoebelen (2012) are examples of building bridges between the two fields, more of which needs to happen in the near future.

In all these developments, it will be important for the field as a whole to avoid fragmentation. The allure of big data should not draw researchers away altogether from ethnographically informed and context-sensitive, qualitative analysis of language use. Many of the discipline's most important achievements would be unthinkable without this latter perspective, and an upsurge in the quantity of data does not in itself guarantee more valid, interesting, or relevant findings. Historically, sociolinguistics has maintained a constant discussion between more qualitatively and more quantitatively oriented researchers. It will be an asset to the field to maintain and intensify this conversation in the context of digital data, methods and questions.

Notes

- 1 The standardisation of corpora involves harmonising the length of corpus samples, applying uniform tag systems to all samples and applying balancing principles in the composition of samples into corpora that take into account language-external features such as the national origin or genre of samples.
- 2 The idea of making a carefully collected and uniformly tagged collection of natural language data available in computer-readable form to researchers is what was first actuated in Brown (containing 500 samples of 2000 words, or 1 million words); it was then adopted in the matching corpora of similar size and genre composition, but representing different national varieties or time periods, LOB, Frown and F-LOB (Hinrichs et al. 2010); and later in numerous corpora reflecting Englishes from around the world.
- 3 To clarify: Szmrecsanyi (2017) does not suggest that the community orientation of sociolinguistics will be replaced by digital studies. We cite him for the presentation of the internal structure of empirical linguistics as a field.
- 4 Machine learning refers to the use of statistical methods in order to develop algorithms that can learn from and make predictions about data (Samuel 1959).
- 5 Some of the messages that users post to Twitter carry location data: they are marked with a precise pair of latitudinal and longitudinal data indicating where the user was at the time of the posting. These data are available, for example, to researchers who download tweets via the Twitter API. Geolocation is provided optionally and only once users have clearly agreed, in their privacy negotiations with the Twitter platform, to the collection and dissemination of this information. About 1 percent of all tweets bear geolocation information.
- 6 Regular expressions allow programmers to formulate searches for text patterns in ways that are more abstract than outright searches for patterns of letters and symbols. A generally known regular expression is the wildcard: symbolised by an asterisk, the use of a wildcard in a corpus search stands for 'any or no intervening material'. So the search string 'is *ing' would retrieve from a corpus any instance of the word *is* followed by any word ending in *-ing*: *is reading*, *is cooking*, *is soliloquizing*, etc. Regular expressions can be formulated at high degrees of abstraction and complication; the interested reader is referred to instruction manuals such as Levithan and Goyvaerts (2012) for a systematic introduction to the topic.
- 7 The corpus text identifier provided with this example is to be read as follows: ICE represents a text from the ICE corpora, the information following after the next dash encodes the national variety represented (here: Irish English), followed by the text category (W2F, i.e. fiction writing) and finally a unique text identifier.

Further reading

- 1 Friedrich, P. and de Figueiredo, E. (2016). *The sociolinguistics of digital Englishes*. London: Routledge. This book is a highly accessible introduction to many theoretical concepts within sociolinguistics, illustrating how these play out in situations of digital communication.
- 2 Georgakopoulou, A. and Spilioti, T. (eds.). (2015). *The Routledge handbook of language and digital communication*. London: Routledge. This edited volume brings together many of the leading researchers in the study of computer-mediated discourse and provides thorough coverage of different methods, research traditions and foci of interest.
- 3 Nguyen, D., Doğruöz, A., Rosé, C. and de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics* 42(3): 537–593. This survey article discusses the potential for synergy and closer collaboration between computational linguistics and sociolinguistics primarily from the perspective of the former discipline, providing countless examples of extant theoretical debates and empirical research.
- 4 Squires, L. (ed.). (2016). *English in computer-mediated communication: Variation, representation, and change*. Berlin: Mouton de Gruyter. Similar to Georgakopoulou and Spilioti (eds., 2015), this edited volume covers the breadth of theoretical, methodological and empirical questions in computer-mediated communication, but more narrowly focused on English language data and situation.

References

- Akkaya, A. (2012). *Devotion and friendship through Facebook: An ethnographic approach to language, community, and identity performances of young Turkish-American women*. PhD, University of Southern Illinois.
- Andersen, G. and Bech, K. (eds.). (2013). *English corpus linguistics: Variation in time, space and genre. Selected papers from ICAME 32*. Amsterdam: Rodopi.
- Androutsopoulos, J.K. (2010). Localizing the global on the participatory web. In N. Coupland (ed.), *The handbook of language and globalization*. Malden, MA: Wiley-Blackwell, pp. 201–231.
- Archer, D. (ed.). (2009). *What's in a word-list? Investigating word frequency and keyword extraction*. London: Routledge.
- Bamman, D., Eisenstein, J. and Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2): 135–160.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Bieswanger, M. (2016). Electronically-mediated Englishes: Synchronicity revisited. In L. Squires (ed.), *English in computer-mediated communication: Variation, representation, and change*. Berlin: Mouton de Gruyter, pp. 281–300.
- Blommaert, J. (2005). *Discourse: A critical introduction*. Cambridge: Cambridge University Press.
- Blommaert, J. (2011). Supervenaculars and their dialects. *Working Papers in Urban Language & Literacies* 81: 1–14.
- Blommaert, J. (2012). Supervenaculars and their dialects. *Dutch Journal of Applied Linguistics* 1(1): 1–14.
- Blommaert, J. (2013). Writing as a sociolinguistic object. *Journal of Sociolinguistics* 17(4): 440–459.
- Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glott International* 5(9–10): 341–345.
- Bohmann, A. (2019). *Variation in English world-wide: Registers and global varieties*. Studies in English Language. Cambridge: Cambridge University Press.
- Bohmann, A. (2016). “Nobody canna cross it”: Language-ideological dimensions of hypercorrect speech in Jamaica. *English Language & Linguistics* 20(1): 129–152.
- Borgman, C. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: MIT Press.

- boyd, danah and Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society* 15(5): 662–679.
- Brugman, H. and Russell, A. (2004). Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of LREC*.
- Callier, P. (2016). Exploring stylistic co-variation on Twitter: The case of DH. In L. Squires (ed.), *English in computer-mediated communication: Variation, representation, and change*. Berlin: Mouton de Gruyter, pp. 241–260.
- Carter, C.J., Koene, A., Perez, E., Statache, R., Adolphs, S., O'Malley, C., Rodden, T. and McAuley, D. (2016). Understanding academic attitudes towards the ethical challenges posed by social media research. *ACM SIGCAS Computers and Society* 45(3): 202–210.
- Cedergren, H. and Sankoff, D. (1974). Variable rules: Performance as a statistical reflection of competence. *Language* 50(2): 333–355.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
- Chun, E. (2013). Ironic blackness as masculine cool: Asian American language and authenticity on YouTube. *Applied Linguistics* 34(5): 592–612.
- Danet, B. and Herring, S.C. (eds.). (2007). *The multilingual Internet: Language, culture, and communication online*. Oxford: Oxford University Press.
- Davies, M. and Fuchs, R. (2015). Expanding horizons in the study of world Englishes with the 1.9 billion word global web-based English corpus (GloWbE). *English World-Wide* 36(1): 1–28.
- Drucker, J. (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly* 5(1). Available at: www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.
- Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology* 41(1): 87–100.
- Eisenstein, J. (2013). What to do about bad language on the Internet. In *Proceedings of the NAACL-HLT 2013*. Atlanta, GA: Association for Computational Linguistics, pp. 359–369.
- Eisenstein, J. (2015). Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics* 19(2): 161–188.
- Fink, C., Kopecky, J. and Morawski, M. (2012). Inferring gender from the content of tweets: A region specific example. In *Proceedings of the sixth international AAAI conference on weblogs and social media*. Palo Alto, CA: AAAI Press, pp. 459–462.
- Georgakopoulou, A. (2011). “On for drinkies?”: Email cues of participant alignments. *Language@Internet* 11. Available at: www.languageatinternet.org/articles/2011/Georgakopoulou.
- Georgakopoulou, A. and Spilioti, T. (eds.). (2015). *The Routledge handbook of language and digital communication*. London: Routledge.
- Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. London: Harper and Row.
- Goffman, E. (1981). *Forms of talk*. Philadelphia: University of Pennsylvania Press.
- Goswami, S., Sarkar, S. and Rustagi, M. (2009). Stylometric analysis of bloggers’ age and gender. In *Proceedings of the third international ICWSM conference*. Menlo Park, CA: AAAI Press, pp. 214–217.
- Greenbaum, S. and Nelson, G. (1996). The international corpus of English (ICE) project. *World Englishes* 15(1): 3–15.
- Grieve, J. (2016). *Regional variation in written American English*. Cambridge: Cambridge University Press.
- Grieve, J., Asnaghi, C. and Ruetten, T. (2013). Site-restricted web searches for data collection in regional dialectology. *American Speech* 88: 413–440.
- Haddican, B. and Johnson, D.E. (2012). Effects on the particle verb alternation across English dialects. *University of Pennsylvania Working Papers in Linguistics* 18(2). Available at: <http://repository.upenn.edu/pwpl/vol18/iss2/5>.
- Heller, M. (2010). *Paths to post-nationalism: A critical ethnography of language and identity*. Oxford: Oxford University Press.
- Herring, S.C. (2002). Computer-mediated communication on the Internet. *Annual Review of Information Science and Technology* 36(1): 109–168.

- Herring, S.C. (2004). Computer-mediated discourse analysis: An approach to researching online behavior. In S.A. Barab, R. Kling and J.H. Gray (eds.), *Designing for virtual communities in the service of learning*. Cambridge: Cambridge University Press, pp. 338–376.
- Herring, S.C. (2007). A faceted classification scheme for computer-mediated discourse. *Language@Internet* 4(1). Available at: www.languageatinternet.org/articles/2007/761.
- Herring, S.C. and Androutsopoulos, J. (2015). Computer-mediated discourse 2.0. In D. Tannen, H. Hamilton and D. Schiffrin (eds.), *The handbook of discourse analysis*, vol. 2. Oxford: Wiley-Blackwell, pp. 127–151.
- Herring, S.C. and Paolillo, J. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics* 10(4): 439–459.
- Heyd, T. (2015). CMC genres and processes of remediation. In A. Georgakopoulou and T. Spiloti (eds.), *The Routledge handbook of language and digital communication*. London: Routledge, pp. 87–102.
- Heyd, T. and Honkanen, M. (2015). From *Naija* to *Chitown*: The new African diaspora and digital representations of place. *Discourse, Context & Media* 9: 14–23.
- Heyd, T. and Mair, C. (2014). From vernacular to digital ethnolinguistic repertoire: The case of Nigerian Pidgin. In V. Lacoste, J. Leimgruber and T. Breyer (eds.), *Indexing authenticity*. Berlin: Mouton de Gruyter, pp. 244–268.
- Hinrichs, L. (2006). *Codeswitching on the web: English and Jamaican Creole in e-mail communication*. Amsterdam: John Benjamins.
- Hinrichs, L. (2018). The language of diasporic blogs: A framework for the study of rhetoricity in written online code-switching. In C. Cutler and U. Rønynland (eds.), *Analyzing youth practices in computer-mediated communication*. Cambridge: Cambridge University Press, pp. 186–204.
- Hinrichs, L., Smith, N. and Waibel, B. (2010). Manual of information for the part-of-speech-tagged, post-edited “Brown” corpora. *ICAME Journal* 34: 189–231.
- Hinrichs, L. and White-Sustaíta, J. (2011). Global Englishes and the sociolinguistics of spelling: A study of Jamaican blog and email writing. *English World-Wide* 32(1): 46–73.
- Huang, Y., Guo, D., Kasakoff, A. and Grieve, J. (2016). Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59: 244–255.
- Iorio, J. (2012). A variationist approach to text: What role-players can teach us about form and meaning. *Texas Studies in Literature and Language* 54(3): 381–401.
- Jaffe, A. (1999). *Ideologies in action: Language politics on Corsica*. Berlin: Mouton de Gruyter.
- Johnson, D.E. (2009). Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3(1): 359–383.
- Johnstone, B. (2013). *Speaking Pittsburghese: The story of a dialect*. Oxford: Oxford University Press.
- Jones, G.M. and Schieffelin, B. (2009). Enquoting voices, accomplishing talk: Uses of *be* + *like* in instant messaging. *Language & Communication* 29(1): 77–113.
- Jones, S., Cotterill, R., Dewdney, N., Muir, K. and Joinson, A. (2014). Finding Zelig in text: A measure for normalising linguistic accommodation. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*. Dublin, Ireland, pp. 455–465.
- Jones, T. (2016). Tweets as graffiti: What the reconstruction of Vulgar Latin can tell us about black Twitter. In L. Squires (ed.), *English in computer-mediated communication: Variation, representation, and change*. Berlin: Mouton de Gruyter, pp. 43–68.
- Keating, E. and Sunakawa, C. (2010). Participation cues: Coordinating activity and collaboration in complex online gaming worlds. *Language in Society* 39(3): 331–356.
- Kendall, T. (2011). Corpora from a sociolinguistic perspective. *Revista Brasileira de Linguística Aplicada* 11(2): 361–389.
- Kiesling, S. (2018). YouTube Yinzers: Stancetaking and the performance of “Pittsburghese”. In R. Bassiouney (ed.), *Identity and dialect performance*. London: Routledge, pp. 245–264.
- Kisler, T., Reichel, U. and Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language* 45: 326–347.
- Kučera, H. and Francis, W.N. (1967). *Computational analysis of present-day English*. Providence: Brown University Press.

- Labov, W. (1966). *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.
- Labov, W. (1984). Field methods of the project in linguistic change and variation. In J. Baugh and J. Sherzer (eds.), *Language in use*. Englewood Cliffs: Prentice Hall, pp. 28–53.
- Levithan, S. and Goyvaerts, J. (2012). *Regular expressions cookbook*. Sebastopol, CA: O'Reilly.
- Licoppe, C. and Morel, J. (2012). Video-in-interaction: “Talking heads” and the multimodal organization of mobile and Skype video calls. *Research on Language and Social Interaction* 45(4): 399–429.
- Mair, C. (2006). Tracking ongoing grammatical change and recent diversification in present-day Standard English: The complementary role of small and large corpora. In A. Renouf and A. Kehoe (eds.), *The changing face of corpus linguistics*. Amsterdam: Rodopi, pp. 355–376.
- Mair, C. (2013). The world system of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide* 34(3): 253–278.
- Mair, C. (2015). Response to Davies and Fuchs. *English World-Wide* 36(1): 29–33.
- Marwick, A. and boyd, d. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society* 13(1): 114–133.
- Mendoza-Denton, N. (2008). *Homegirls: Language and cultural practice among Latina youth gangs*. Malden, MA: Blackwell.
- Milroy, L. (1980). *Language and social networks*. Baltimore: University Park Press.
- Moisl, H. (2015). *Cluster analysis for corpus linguistics*. Berlin: Mouton de Gruyter.
- Nelson, G. (2015). Response to Davies and Fuchs. *English World-Wide* 36(1): 38–40.
- Nerbonne, J. and Kretzschmar, W. (2003). Introducing computational techniques in dialectometry. *Computers and the Humanities* 37(3): 245–255.
- Nguyen, D., Doğruöz, A., Rosé, C. and de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics* 42(3): 537–593.
- Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. London: Viking.
- Patrick, P. (1999). *Urban Jamaican Creole: Variation in the mesolect*. Amsterdam: John Benjamins.
- Pavalanathan, U. and Eisenstein, J. (2015). Audience-modulated variation in online social media. *American Speech* 90(2): 187–213.
- Pennycook, A. (2003). Global Englishes, rip slyme, and performativity. *Journal of Sociolinguistics* 7(4): 513–533.
- Peuronen, S. (2011). “Ride hard, live forever”: Translocal identities in an online community of extreme sports Christians. In C. Thurlow and K. Mroczek (eds.), *Digital discourse: Language in the new media*. Oxford: Oxford University Press, pp. 154–176.
- Posner, M. (2012). Think talk make do: Power and the digital humanities. *Journal of Digital Humanities* 1(2). Available at: <http://journalofdigitalhumanities.org/1-2/think-talk-make-do-power-and-the-digital-humanities-by-miriam-posner>.
- Reddy, S. and Stanford, J. (2015). Toward completely automated vowel extraction: Introducing DARLA. *Linguistics Vanguard* 1(1): 15–28.
- Rintel, E.S., Mulholland, J. and Pittam, J. (2001). First things first: Internet relay chat openings. *Journal of Computer-Mediated Communication* 6(3). <https://doi.org/10.1111/j.1083-6101.2001.tb00125.x>.
- Rosenfelder, I., Fruehwald, J., Evanini, K. and Yuan, J. (2011). *FAVE (Forced Alignment and Vowel Extraction) Program suite*. Available at: <http://fave.ling.upenn.edu/>.
- Sacks, H., Schegloff, E.A. and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50(4): 696–735.
- Samuel, A. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* 3(3): 210–229.
- Sankoff, G. and Cedergren, H. (1972). Sociolinguistic research on French in Montreal. *Language in Society* 1(1): 173–174.
- Schnoebelen, T. (2012). *Emotions are relational: Positioning and the use of affective linguistic resources*. PhD, Stanford University.

- Schöch, C. (2013). Big? Smart? Clean? Messy? Data in the humanities. *Journal of Digital Humanities* 2(3). Available at: <http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities>.
- Schreier, D. (2008). *St Helenian English: Origins, evolution and variation*. Amsterdam: John Benjamins.
- Seargeant, P. and Tagg, C. (2011). English on the Internet and a “post-varieties” approach to language. *World Englishes* 30(4): 496–514.
- Smakman, D. (2015). The westernizing mechanisms in sociolinguistics. In D. Smakman and P. Heinrich (eds.), *Globalising sociolinguistics: Challenging and expanding theory*. London: Routledge, pp. 16–35.
- Spilioti, T. (2011). Beyond genre: Closings and relational work in text-messaging. In C. Thurlow and K. Mroczek (eds.), *Digital discourse: Language in the new media*. Oxford: Oxford University Press, pp. 67–85.
- Squires, L. (2010). Enregistering Internet language. *Language in Society* 39(4): 457–492.
- Squires, L. (2014). From TV personality to fans and beyond: Indexical bleaching and the diffusion of a media innovation. *Journal of Linguistic Anthropology* 24(1): 42–62.
- Squires, L. (ed.). (2016). *English in computer-mediated communication: Variation, representation, and change*. Berlin: Mouton de Gruyter.
- Squires, L. and Iorio, J. (2014). Tweets in the news: Legitimizing medium, standardizing form. In J. Androutsopoulos (ed.), *Mediatization and sociolinguistic change*. Berlin: Mouton de Gruyter, pp. 331–360.
- Stommel, W. (2008). Conversation analysis and community of practice as approaches to studying online community. *Language@Internet*: 5. Available at: www.languageatinternet.org/articles/2008/1537.
- Szmrecsanyi, B. (2013). *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge: Cambridge University Press.
- Szmrecsanyi, B. (2017). Variationist sociolinguistics and corpus-based variationist linguistics: Overlap and cross-pollination potential. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 62(4): 685–701.
- Tagliamonte, S. and Baayen, R.H. (2012). Models, forests and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(2): 135–178.
- Tagliamonte, S. and Denis, D. (2008). Linguistic ruin? Lol! Instant messaging and teen language. *American Speech* 83(1): 3–34.
- Tamminga, M., Ahern, C. and Ecay, A. (2016). Generalized additive mixed models for intraspeaker variation. *Linguistics Vanguard* 2(1): 1–9.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.
- Thurlow, C. (2006). From statistical panic to moral panic: The metadiscursive construction and popular exaggeration of new media language in the print media. *Journal of Computer-Mediated Communication* 11(3): 667–701.
- Thurlow, C. and Mroczek, K. (eds.). (2011). *Digital discourse: Language in the new media*. Oxford: Oxford University Press.
- Vandergriff, I. (2014). A pragmatic investigation of emoticon use in nonnative/native speaker text chat. *Language@Internet*: 11. Available at: www.languageatinternet.org/articles/2014/vandergriff.
- Weber, H.L. (2011). Missed cues: How disputes can socialize virtual newcomers. *Language@Internet*: 9. Available at: www.languageatinternet.org/articles/2011/Weber.
- Weinreich, U., Labov, W. and Herzog, M.I. (1968). Empirical foundations for a theory of language change. In W.P. Lehmann and Y. Malkiel (eds.), *Directions for historical linguistics*. Austin, TX: University of Texas Press, pp. 95–192.
- Wiebesiek, L., Rudwick, S. and Zeller, J. (2011). South African Indian English: A qualitative study of attitudes. *World Englishes* 30(2): 251–268.
- Wunder, E., Voormann, H. and Gut, U. (2010). The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. *ICAME Journal* 34: 78–88.

- Yates, S. (1996). Oral and written linguistic aspects of computer conferencing. In S.C. Herring (ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives*. Amsterdam: John Benjamins, pp. 29–46.
- Zappavigna, M. (2015). Searchable talk: The linguistic functions of hashtags. *Social Semiotics* 25(3): 274–291.

Literary stylistics

Michaela Mahlberg and Viola Wiegand

Introduction

Literary stylistics is a discipline that integrates the study of language and literature. Wales (2001: 371) defines ‘style’ as ‘the set or sum of linguistic features that seem to be characteristic: whether of register, genre or period, etc.’. Literary stylistics is an applied discipline that draws on a range of linguistic approaches, models and methods to study literary texts. Stockwell and Whiteley (2014a: 1) go as far as to see stylistics as ‘the proper study of literature’. By calling it ‘proper’, they emphasise that stylistic analysis is systematic and closely engages with the actual text. Instead of ‘literary stylistics’ the term ‘literary linguistics’ is also used, and ‘stylistics’ is not automatically seen as dealing with literary texts, but can refer to the study of style more widely. In this chapter, we approach stylistics as the study of textual features and their functions in literary texts, with a particular focus on prose fiction.

In the digital age, stylistics has benefitted from the availability of digitised versions of literary texts and methods of textual analysis that are supported by simple visualisation techniques, as well as sophisticated quantitative approaches. The term ‘corpus stylistics’ has come to be used to refer to research in this area. In the present chapter, we deal with the contribution that corpus methods can make to the study of literary texts. The chapter focuses on similarities between analytical approaches to narrative fiction in corpus linguistics and digital humanities. In this sense, the chapter brings together three areas of research – literary stylistics, corpus linguistics and digital humanities – that theoretically have much in common, but in practice more often than not operate within disciplinary boundaries. We argue that stylistics provides a theoretical foundation for the analysis of texts in the digital age that is beneficial for both corpus linguistics and the wider digital humanities.

Background

To some extent, trends in stylistics can be seen as a reflection of developments in linguistics. Accordingly, stylistics has been shaped by formalist, functionalist and pragmatic traditions in the twentieth century. For detailed accounts of the development and evolution of stylistics more generally, see for instance Nørgaard et al. (2010), Stockwell (2006), Wales (2014),

Zyngier and Fialho (2015). As we will show in the ‘Current contributions and research’ section, more recently, cognitive and corpus linguistic approaches have shaped important developments in stylistics. Simpson (2014: 3) describes stylistics as ‘a method of textual interpretation in which primacy of place is assigned to *language*’. With this focus on language, we have the ‘full array of language models at our disposal’ (Simpson 2014: 3–4), and new developments in linguistics can have an effect on how stylistics is done.

Irrespective of the trends and fashions in linguistics, however, literary stylistics is characterised by an interest in the literary qualities and aesthetic purposes of a text. In stylistics, the term ‘foregrounding’ is often used to account for linguistic features that are made prominent in literary works and stand out against the background of a specific textual context or deviate from the norms of the language in general. Foregrounding is a psychological phenomenon. It is the effect brought about in the reader when deviant features of a text are made perceptually prominent (cf. Leech 1985: 47; Short 1996: 11). Emmott and Alexander (2014: 330) make it clear that ‘foregrounding’ is not limited to ‘literary’ concerns. It refers more generally to any example of language that may receive attention, so it ‘relates simply to whether an item is likely to be noticeable or not’. Still, foregrounding plays a crucial role for literary stylistics because a literary text can be read both as a sample of language and as a work of art. With reference to Spitzer’s (1948) ‘philological circle’, Leech and Short (2007: 11) describe the aim of literary stylistics as ‘to relate the critic’s concern of aesthetic appreciation with the linguist’s concern of linguistic description’.

Today, stylistics is a thriving discipline in its own right. The maturity of the field was particularly put into focus when three major handbooks were published within the space of two years: *The Routledge Handbook of Stylistics* edited by Burke (2014), *The Cambridge Handbook of Stylistics* by Stockwell and Whiteley (2014b) and *The Bloomsbury Companion to Stylistics* by Sotirova (2015). In her review of the three reference works, Sorlin (2016: 287) sees it as testament to the ‘vitality’ of the discipline that the handbooks are not competing for the same space but are able to provide a complementary picture of an expanding field. The expansion and especially the interdisciplinarity that is key to literary stylistics is crucial for the view of literary stylistics in the wider context of the digital humanities that we will present in this chapter.

Critical issues and topics

Especially for introductory purposes, stylistics is often presented with a ‘toolkit’ metaphor (cf. Leech and Short 2007), which is sometimes further facilitated with ‘checklists’ and ‘checksheets’ (Short 1996; Stockwell 2010). However, Wales (2014: 32) points out that the checklist view may create the false impression that a successful analysis requires ‘ticking [all] the boxes’. She also warns that ‘[w]herever the stylistician may roam, the text must be the starting and finishing point, with the interpretation as the goal’ (Wales 2014: 45).

With the notion of ‘stylistics as a toolkit’ often comes an emphasis on the methods that are available for textual analysis. Toolan (2014: 13), however, stresses that

stylistics does have a theory – or . . . different forms of stylistics reflect different underlying theories, clearly articulated or otherwise. In large degree the theories stylisticians at least tacitly invoke are theories of language, which they inherit from the particular kind of linguistics (systemic-functional, corpus, cognitive, etc.) they chiefly employ.

In this sense, literary stylistics is similar to corpus linguistics, where debates about the state of the field in terms of theory or method are not uncommon; see e.g. Pope (2010).

Toolan (2014: 16–20) further presents a list of 11 ‘facts’ that he argues most stylisticians base their work on. Here we summarise those that are most relevant to our argument in the present chapter (the numbering is not Toolan 2014’s original numbering):

- 1 ‘Language is central to the meaning . . . of texts’.
- 2 Given the important cultural status of literature, it is justifiable to pay close attention to the forms and functions of its language features.
- 3 The linguistic forms in a literary work are chosen for a reason; i.e. every feature is meaningful.
- 4 Meaning does not only derive from the text; the analysis focuses on the text but considers contextual factors.
- 5 Texts are mostly selected for analysis because they are ‘highly valued and highly crafted’; examples are chosen ‘to advance our understanding of literary linguistic phenomena’ and stylistic work will never end, as receptions change and new methods emerge.
- 6 The analysis of an individual work can never be exhaustive; the level of detail depends on the aims of the study.
- 7 Stylistics provides great value to the student of literature due to its transparent explanations and it can offer new perspectives on the analysed work as well as on reading, writing and language more generally.
- 8 Considered in terms of its individual linguistic features, a literary text operates like any other type of language, but, ‘taken as a whole, remain[s] much more intellectually interesting’.

Toolan’s (2014) points focus mainly on the study of literature. The last point, in particular, creates a value distinction between literary and non-literary texts. Toolan (2014: 19) observes that ‘the literary text is different, and aims to be’, having been crafted with particular skill. This argument seemingly depicts literariness as an ‘either or quality’. However, corpus linguistic work contributes to a view of literariness along a cline (cf. Carter 2004). Moreover, not all stylistic research focuses on literature. A number of stylisticians have been working with non-fiction: the research programme of ‘critical stylistics’ specifically takes stylistic approaches closer to discourse analysis, as becomes apparent in the analysis of texts like newspaper articles (e.g. Jeffries and Walker 2018).

The boundaries of the discipline are not only concerned with the types of texts under investigation but also with the methodological approaches that are employed. Stylometry is closely related to stylistics in that both are concerned with style. However, objectives and methods of researchers in the two camps differ. Stylistics places much emphasis on the transparent and systematic analysis of linguistic features in the text(s) under investigation. Such analyses may range from the detailed study of a single book (see e.g. Giovanelli 2018 for a recent example exploring mind style in Paula Hawkins’s *The Girl on the Train*) to larger-scale analyses across many texts with the help of corpus methods (e.g. Mahlberg 2013; Semino and Short 2004). As Toolan’s (2014) points indicate, stylistics has an inherent interest in meaning as part of its focus on style. By contrast, stylometric approaches tend to be less concerned with meaning-making patterns and more with the identification and comparison of stylistic features that are relevant, for instance, to address questions of authorship (e.g. Burrows 2002; Eder et al. 2016). In this sense, stylometry takes the ‘distinctive narrative style’ outlined by Wales (2014: 40) as a feature to distinguish between authors.

While quantitative approaches seem to be becoming increasingly popular in the digital age, fundamental principles of quantification have been around for a long time – and even before the advent of computers. Reviewing the history of ‘humanities computing’ as part of the *Companion to Digital Humanities*, Hockey (2004) provides an account of a late-nineteenth-century method of manually counting stylistic features such as sentence length in order to identify authorship. Hockey (2004) refers to Mendenhall (1901) as an example of the application of this method. This early work already draws on the notion of ‘unconscious’ choices, such as sentence length and function words, that make up an author’s style and that are still central to modern stylometric and authorship analysis (see e.g. Hoover 2007: 176).

The digital humanities, stylometry and corpus stylistics share an interest in digital approaches to text analysis. In his essay on the purpose of the digital humanities in English departments, Kirschenbaum (2010: 60) points out that

after numeric input, text has been by far the most tractable data type for computers to manipulate. Unlike images, audio, video, and so on, there is a long tradition of text-based data processing that was within the capabilities of even some of the earliest computer systems and that has for decades fed research in fields like stylistics, linguistics, and author attribution studies, all heavily associated with English departments.

Important technological innovations have been developed within the digital humanities. One of them is the formation of the Extensible Markup Language (XML) standards of the Text Encoding Initiative (Burnard 2014; Hockey 2000; Terras 2010, 2016), which have also been applied in corpus linguistics. However, it seems that obvious ties between the disciplinary camps have not been leveraged enough. Jockers (2013: 18) sees one of the reasons why corpus linguists and digital literary scholars have taken different directions in the fact that corpus linguists have been developing specific resources, such as the British National Corpus (BNC), and stand-alone tools to understand how language works, while literary scholars ‘have generally been content to rely on the web for access to digital material’. Already in the early 2000s, Hockey (2004: online) comments that ‘the loss of some aspects of linguistic computing, particularly corpus linguistics, to conferences and meetings of its own’ in the 1980s was an unfortunate event for the community that has since grown into the digital humanities. She continues that ‘there was little communication between these communities, and humanities computing did not benefit as it could have done from computational linguistics techniques’ (Hockey 2004: online).

The range of tools that have been developed in the digital humanities community include initiatives aiming to enable a wider audience to analyse text digitally. For example, the iOS app Textal (Terras et al. 2013) was created to easily generate word counts and word clouds of texts from websites and Twitter streams using a phone. The Computer Assisted Textual Markup and Analysis (CATMA) web application (Meister et al. 2019) seems to be particularly designed for the annotation of text to support analysis, but it also contains options for generating word lists and KWIC displays, that is, concordances (see the ‘Current contributions and research’ section). Another web interface, Voyant Tools (S. Sinclair and Rockwell 2018, n.d.), brings together various digital text analysis tools with a particular focus on visual outputs. Again, this interface includes concordances, but it also features more idiosyncratic ‘conceptual’ plots like the ‘Knots’ tool that represents occurrences of a word in a text as a twisted line (the more twisted, the more frequent the word) or the ‘Mandala’ tool that visualises relations between terms and various documents with colourful lines. An example of the application of Voyant Tools is provided in S. Sinclair and Rockwell (2015).

There are clear similarities between such digital humanities tools and those used in corpus linguistics: both communities use tools to count words, explore their usage across texts and visualise the results. If created effectively and accurately, visualisations can help to summarise the complex relationships in large-scale textual data and therefore aid the analysis. Anthony (2018) discusses some of the best practices and pitfalls of visualisation in a chapter aimed at corpus linguists, encouraging them to use a wider range of visualisation techniques. In a similar vein, but targeting a different audience, a comparison of various network and cluster visualisation techniques is given by Eder (2017: 62), who aims to ‘encourage stylometrists to produce a reliable map of literature in its entirety, and to propose a methodological background for such a map’.

Thematically, an example of a potentially shared interest between the digital humanities and literary stylistics is the analysis of fictional characters. In literary stylistics, the study of characterisation has benefitted from corpus approaches (e.g. Culpeper 2009; Stockwell and Mahlberg 2015). An example of a digital humanities approach to fictional characters is Galleron (2017) who makes the case for creating a systematic classification of characters in French plays from 1630 to 1810. Unlike corpus stylistic approaches, this project (see also Galleron et al. 2018) aimed to recruit contributors for ‘crowd-reading’ the plays and encoding the character information in XML versions of the text, thereby creating a digital resource for future study. Despite the common interests, as Jensen (2014: 115) points out, corpus linguistics ‘remains rather obscure in most contemporary literature on the digital humanities’. As the digital humanities develop, questions around ‘inclusion and exclusion’ receive attention (Nyhan et al. 2013: 3), that is, which researchers and approaches ‘belong’ to the community. At the same time, researchers like Terras (2016: 1644) stress the need to keep reflecting on how the tools that are developed relate to and impact the humanities.

As shown earlier, inward-looking reflections on disciplinary developments have an important place in the digital humanities, stylistics and corpus linguistics, respectively. In this chapter, we argue that it is equally important to take an outward-looking perspective that enables greater dialogue between the different research communities.

Current contributions and research

In the preface to the second edition of his stylistics textbook, Paul Simpson (2014: xviii) reflects on how changes within the field, since the publication of the first edition, have shaped the modification of the textbook: ‘[t]en years is a long time in stylistics . . . this second edition is in its scope a bit more multimodal, a bit more cognitive and a lot more corpus-assisted’. Similarly, Leech and Short (2007: 286) refer to a ‘cognitive turn’ in stylistics that saw concepts and methods from cognitive linguistics and psychology lead to the thriving area of cognitive stylistics or cognitive poetics (see e.g. Gavins and Steen 2003; Semino and Culpeper 2002; Stockwell 2002). This line of research places particular emphasis on the readerly effects of texts; going back to point (6) of Toolan’s (2014) fundamental stylistic tenets listed earlier, cognitive poetics provides (some of the) elements of meaning that are not inherent in the text. Across a similar time period, the use of corpus linguistic methods has had considerable impact on stylistics and can be seen as a ‘corpus turn’ (Leech and Short 2007: 286) leading to ‘corpus stylistic’ approaches.

While Hockey (2004) observes that the growth of corpus linguistics in its own right resulted in a lack of communication with the digital humanities community, we argue that the development of greater specialism within corpus linguistics in the form of corpus stylistics opens up new opportunities for corpus linguistics and digital humanities to reach out

to one another. There are obvious resemblances between corpus stylistics and approaches in the digital humanities. Jockers (2013: 32) proposes the method of ‘macroanalysis’ to study trends and patterns in ‘the digital-textual world’. Lkening macroanalysis to macroeconomics, he argues that ‘the study of literature should be approached not simply as an examination of seminal works but as an examination of an aggregated ecosystem or “economy” of texts’ which makes it possible to contextualise individual works (Jockers 2013: 32). An example is his study of thematic and topical patterns across a corpus of 3346 nineteenth-century books that reveal gendered patterns of topics across the authors in the corpus. Similarly, a study by Underwood et al. (2018) on how the language used to describe fictional men and women has changed since the eighteenth century highlights that gender seems to be a particularly revealing parameter for the study of literature at scale. The emphasis on trends and the bigger picture is also central to what characterises approaches of ‘distant reading’ (Moretti 2013; Underwood 2017; Klein 2018), which are reminiscent of John Sinclair’s (1991: 100) famous observation that ‘[t]he language looks rather different when you look at a lot of it at once’.

J. Sinclair’s (1991) observation is fundamental to research in corpus linguistics and in particular to the corpus linguistic use of ‘concordances’. A concordance is a display format where the ‘node’, the word that the researcher is interested in, is displayed in the centre with a specified amount of context on the left and on the right (cf. the examples in the sample analysis below). Concordances are a key means by which researchers identify patterns of words and the meanings that are associated with these patterns. J. Sinclair (2003) explains this fundamental method in his textbook pointedly entitled *Reading Concordances*. Crucial to the reading of concordances is that patterns are identified by reading ‘vertically’, scanning the lines up and down rather than reading each line purely ‘horizontally’ as we would read a text (cf. Tognini-Bonelli 2001). As such, the concordance is a basic format of visualisation, even though it is highly text-based. In a later section, we will introduce the CLiC web app, a concordance tool specifically designed for the study of literary texts, as it makes it possible to run concordances of subsections of the text that are relevant for stylistic analysis (such as character speech) and group the retrieved lines according to textual patterns.

While the reading of concordances increases the distance from individual texts, it is not yet as distant as most approaches of ‘distant reading’. Crucially, however, corpus linguistics and digital humanities seem to have so far missed the opportunity to bring their approaches to reading together. Whereas in corpus linguistics concordance reading is a widely shared method, in digital humanities, it seems the concept of ‘reading’ is one way of developing variety in approaches. Best and Marcus (2009) use the term ‘surface reading’ of literary text to contrast the activity with what they view as mainly interpretive ‘symptomatic reading’ of literary critics (Best and Marcus 2009: 4). Lee et al. (2018) propose a method that they call ‘Linked Reading’ which aims to combine the analysis of different datasets. The authors illustrate the approach with a study of discourses of race in Shakespeare’s *Othello*. They use topic and vector space models to mine texts from the Early English Books Online (EEBO) corpus and combine linguistic analysis with the study of historical book networks of publishers, printers and booksellers. There is also terminological variety that characterises approaches in the wider digital humanities without explicit reference to reading. Terms such as ‘cultural analytics’ or ‘culturomics’ have been proposed for the analysis of large amounts of textual data, which does not always have to be literary texts. While such terminology can stress the novelty of technological developments and the amount and scale of the data under investigation, it also indicates that researchers other than literary scholars or linguists are interested in the analysis of (literary) texts, which can lead to differing views on

the ways in which to interpret the data. A prominent example is the study of over 5 million Google Books titles by Michel et al. (2011: 176), claiming to ‘survey the vast terrain of “culturomics”, focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000’. Pechenick et al. (2015) argue that such endeavours are rather ambitious. They point out statistical limitations of making sociocultural inferences from trends in Google Books due to the unbalanced nature of the Google Books corpus.

We started this chapter by linking the analysis of language and literature. It is useful to return to this link to view literary stylistics in the context of the digital humanities. Hall (2014: 99) makes a strong point for literary critics and stylisticians benefitting from each other’s work:

Stylisticians must learn to read and respect the contributions the expertise of the literary critics can bring to discussions of literary works; literary critics must be shown the value of getting their hands dirty with a little practical linguistic analysis.

A similar argument applies to the computer-assisted analysis of literary texts where scholars from corpus linguistics and digital humanities can come together. In particular, we see the place of corpus stylistics as a mediator between stylistic approaches that focus on qualities of individual texts while drawing on linguistic frameworks and larger-scale quantitative endeavours that are typically considered as digital humanities. At a time when opportunities of technological advances seem limitless for mining large quantities of text (as in the example of the Google Books corpus), the identification and understanding of shared ground between the various disciplines seems a good way to avoid merely opportunistic description.

Main research methods

Any linguistic framework can be adopted for stylistic analysis. Additionally, stylistics has taken on methods from other disciplines such as psychology, particularly for more experimental research methods (e.g. Mahlberg et al. 2014) and in the area of reader-response research (for an overview of methods see Whiteley and Canning 2017). In this chapter, our focus is on a corpus stylistic approach. In theory, corpus methods enable us to examine any type of linguistic information. However, some linguistic features lend themselves more easily to corpus linguistic investigations than others. Often, corpus stylistic analyses focus on lexical patterns. The identification of such patterns tends to rely on word-based methods like concordancing (see the previous section), the retrieval of clusters (also called *n*-grams), that is, repeated sequences of words such *I don’t know what*, or the generation of key words. Key words result from the comparison of texts: the frequencies of words in text (or corpus) A are compared with the frequencies of words in corpus B to find the words that occur significantly more frequently in A (i.e. those that are ‘key’). An example is Culpeper’s (2009) comparative study of the speech by individual characters in *Romeo and Juliet*.

If corpora are pre-annotated (e.g. for grammatical information such as parts of speech or semantic meaning of words and phrases) the additional information can add further dimensions to an analysis. Unsurprisingly, stylistic questions, with a different objective from traditional corpus linguistic enquiries, have given rise to new ideas for annotating corpora. Fictional dialogue has been a particularly influential area in this regard. For example, Semino and Short (2004) work with a corpus of prose fiction, news reports and biographies that has been annotated for various classes of discourse presentation including direct speech and more complex cases like free indirect speech (for discourse presentation especially in

nineteenth-century fiction see Busse 2010). Some approaches have moved from lexical to semantic or conceptual patterns. In addition to a key words comparison, Culpeper (2009) compares key semantic domains, that is, the most overused conceptual domains for a given character. Similarly, in a study of Virginia Woolf's *The Waves*, Balossi (2014) makes use of the novel's narrative structure to isolate and compare characters' soliloquies in terms of 'key' semantic domains (e.g. 'geographical names'). Semantic comparisons are not only useful for analysing fictional speech; Mahlberg and McIntyre (2011) compare lexical key words as well as key semantic domains in the James Bond novel *Casino Royale* against the fictional prose section of the BNC. They identify elements that describe the fictional world, including the characters, but also settings (e.g. places and food) and thematic signals.

Conceptual patterns can also be identified based on the categories of a theoretical framework. An example of this approach is the recently developed WorldBuilder tool (Wang et al. 2016), which facilitates the annotation of a text with the elements of the cognitive stylistic framework of text world theory (Gavins 2007; Werth 1999). The WorldBuilder enables the user to quantify text world categories and create 'cognitive diagrams'. This is another example of the visualisation of textual data, in this case, based on a cognitive linguistic theoretical framework rather than frequency features of words in the text. Ho et al. (2018) illustrate the approach with a case study drawing on a statement from the Meredith Kercher murder trial.

As a tool specifically developed for literary stylistic analysis, the CLiC web app (Corpus Linguistics in Context; <http://clic.bham.ac.uk>) has recently made available a collection of fiction corpora containing over 16 million words that have been semi-automatically tagged for character speech and narration (Mahlberg et al. 2016). In addition to standard corpus tool functionalities, CLiC allows the user to restrict searches to text within or outside of quotation marks. Concordance searches can be further refined through 'KWICGrouping' (cf. also O'Donnell 2008) of results. CLiC version 2.0 contains five corpora (the analysis in this chapter was carried out with CLiC version 2.0.0-beta7). In this chapter, we will make use of the corpus of Dickens's novels (DNov; 15 novels; 3.8 million words), the 19th Century Reference Corpus (19C; 29 novels; 4.5 million words) and the 19th Century Children's Literature Corpus (ChiLit; 71 books; 4.4 million words).

Sample analysis

As we outlined earlier, it is crucial for a literary stylistic analysis to combine both literary and linguistic concerns. The handbooks we mention in the 'Background' section point to many excellent examples of such analyses. The aim of this section is not to provide a detailed example of a specific text, but to bring out three types of key contributions that corpus methods can make to the analysis of literary texts. We will illustrate how corpus methods can

- 1 support the analysis of meaning that is created through repetition;
- 2 contribute to the understanding of conventional patterns in narrative;
- 3 contextualise specific literary examples within general language use.

Meaning through repetition

Concordance analysis, a central method in corpus linguistics, shows how repeated occurrences of words provide evidence of linguistic patterns. Such patterns are associated with meanings. So, for corpus linguistics, the relationship between meaning and repetition is

fundamental to the discipline. Repetition also plays a key role in the analysis of literary texts (e.g. Leech and Short 2007; Leech 2008; Wales 2001). It is one of the central forms of linguistic deviation that creates foregrounding effects. To illustrate the stylistic effects of repetition, the opening chapter of Dickens's *Bleak House* often serves as an illustration – '[m]uch beloved of stylisticians because of its foregrounded patterns of language' (Simpson 2014: 65), see also Carter (2010) and Jeffries (1996). One of the features that is typically commented on is the repetition of the word *fog*, specifically the parallel structure in the second paragraph of the chapter with several sentences starting with this word. The 'fog passage' is also discussed in literary criticism, for example, in relation to the theme 'Dickens and the city' (e.g. Mee 2010; Schwarzbach 1990).

While analyses of the opening chapter of *Bleak House* stress the importance of fog for the novel, what is discussed less often is how the word *fog* is referred to in subsequent chapters. In Concordance 17.1, the 'in bk.' display shows that most instances of *fog* appear towards the beginning of the book (lines 1–22 occur in Chapter 1). While the word is repeated later on in the novel, it is much less frequent. Example (1) illustrates the co-text of line 23, where, in Chapter 3, the fog is introduced to Esther as 'a London particular'.

- (1) He was very obliging, and as he handed me into a fly after superintending the removal of my boxes, I asked him whether there was a great fire anywhere? For the streets were so full of dense brown smoke that scarcely anything was to be seen.
 "Oh, dear no, miss," he said. "This is a London particular."
 I had never heard of such a thing.
 "A *fog*, miss," said the young gentleman.
 "Oh, indeed!" said I.

(*Bleak House*, Charles Dickens, Chapter 3)

There are further occurrences in Chapters 4 and 5, but as the novel progresses the references to fog become fewer. *Fog* occurs again in Chapters 17 and 22 (once, respectively) and twice in Chapter 45 (see lines 31 and 32). In Chapter 45, however, it is not the same dense fog from the earlier chapters. The *white fog* that then rises *like a curtain* occurs at the seaside and not in the city. The example of *fog* in *Bleak House* shows how a concordance display can support an analysis by tracing the repetitions of a word throughout the text. The concordance also provides parts of the immediate textual context linked to the meaning in the text. In Concordance 17.1, lines 2 to 5, for instance, contain prepositional phrases and lines 6 to 8 provide examples of non-finite verb forms that all contribute to the dense, static picture of *fog* in Chapter 1. The prominence of *fog* in the opening chapter suggests that fog is a salient presence for the story and contributes to the atmosphere throughout. However, once a specific meaning is created, it might not need explicit verbatim repetition, as evidenced by the actual distribution of the word *fog*. This example also highlights that the meaning of a word is shaped by where in a text it occurs. For a novel, the opening chapter is an important place where the creation of the fictional world starts. While features of this world may be specifically reinforced throughout the text, they might not need repeating. At the same time, the examples from Chapter 45 indicate that a word can take on different meanings depending on the context in which it occurs and the development of the story.

While concordances can track the occurrences of words throughout a novel, in corpus linguistics, concordances are more typically used to identify co-occurrence patterns of words, or 'collocations', that is, repeated occurrences of normally two words within a specified amount of context. Network visualisations have recently become increasingly popular

Left	Node	Right	Book	In bk.
1. Closely to the pavement, and accumulating at compound interest. ¶	Fog	everywhere. Fog up the river, where it flows among green	BH	
2. pavement, and accumulating at compound interest. ¶ Fog everywhere.	Fog	up the river, where it flows among green aits and	BH	
3. the river, where it flows among green aits and meadows;	fog	down the river, where it rolls delfied among the tiers	BH	
4. and the waterside pollutions of a great (and dirty) city.	Fog	on the Essex marshes, fog on the Kentish heights. Fog	BH	
5. a great (and dirty) city. Fog on the Essex marshes,	fog	on the Kentish heights. Fog creeping into the cabooses of	BH	
6. Fog on the Essex marshes, fog on the Kentish heights.	Fog	creeping into the cabooses of collier-brigs; fog lying out on	BH	
7. the Kentish heights. Fog creeping into the cabooses of collier-brigs;	fog	lying out on the yards and hovering in the rigging	BH	
8. the yards and hovering in the rigging of great ships;	fog	drooping on the gunwales of barges and small boats. Fog	BH	
9. fog drooping on the gunwales of barges and small boats.	Fog	in the eyes and throats of ancient Greenwich pensioners, wheezing	BH	
10. ant Greenwich pensioners, wheezing by the firesides of their wards;	fog	in the stem and bowl of the afternoon pipe of	BH	
11. pipe of the watchful skipper, down in his close cabin;	fog	cruelly pinching the toes and fingers of his shivering little	BH	
12. bridges peeping over the parapets into a nether sky of	fog,	with fog all round them, as if they were up	BH	
13. over the parapets into a nether sky of fog, with	fog	all round them, as if they were up in a	BH	
14. and hanging in the misty clouds. ¶ Gas looming through the	fog	in divers places in the streets, much as the sun	BH	
15. unwilling look. ¶ The raw afternoon is rawest, and the dense	fog	is densest, and the muddy streets are muddiest near that	BH	
16. in Lincoln's Inn Hall, at the very heart of the	fog,	sits the Lord High Chancellor in his High Court of	BH	
17. in his High Court of Chancery. ¶ Never can there come	fog	too thick, never can there come mud and mire too	BH	
18. lantern in the roof, where he can see nothing but	fog.	On such an afternoon some score of members of the	BH	
19. dim, with wasting candles here and there; well may the	fog	hang heavy in it, as if it would never get	BH	
20. midst of the mud and at the heart of the	fog,	sits the Lord High Chancellor in his High Court of	BH	
21. voice arises, fully inflated, in the back settlements of the	fog,	and says, "Will your lordship allow me? I appear for	BH	
22. of the roof, the very little counsel drops, and the	fog	knows him no more. Everybody looks for him. Nobody can	BH	
23. particular." ¶ I had never heard of such a thing. ¶ "A	fog,	miss," said the young gentleman. ¶ "Oh, indeed!" said I. ¶ We	BH	
24. moment to ask a question and left us in the	fog,	with the Lord Chancellor's carriage and servants waiting for him	BH	
25. He seemed quite delighted with it on my account. ¶ "The	fog	is very dense indeed!" said I. ¶ "Not that it affects	BH	
26. of high houses like an oblong cistern to hold the	fog.	There was a confused little crowd of people, principally children	BH	
27. no one. ¶ The purblind day was feebly struggling with the	fog	when I opened my eyes to encounter those of a	BH	
28. rning Adventure ¶ Although the morning was raw, and although the	fog	still seemed heavy--I say seemed, for the windows were	BH	
29. trusting when I first saw it in that memorable November	fog,	how much more did it seem now when I knew	BH	
30. se may not be desirable characteristics when November comes with	fog	and sleet or January with ice and snow, but they	BH	
31. ever saw. The sea was heaving under a thick white	fog;	and nothing else was moving but a few early ropemakers	BH	
32. ship's cabin, and that delighted Charley very much. Then the	fog	began to rise like a curtain, and numbers of ships	BH	

Concordance 17.1 All 32 instances of *fog* in *Bleak House*, retrieved with CLiC

to visualise such collocational relationships. The corpus tool #LancsBox (<http://corpora.lancs.ac.uk/lancsbox>) contains a specific module, called GraphColl, for displaying collocation networks, explained in Brezina et al. (2015).

Similar tools to display networks are used by the digital humanities community. In Voyant Tools, for instance, the 'Links' function displays collocates. The top panel in Figure 17.1 shows a Links plot for *fog* in *Bleak House* with relatively frequent collocates within a span of five words to the left and right. In the bottom panel of Figure 17.1, we used the CLiC KWICGrouper to highlight these collocates in their context of a concordance display. As there are only 32 occurrences of *fog* (see Concordance 17.1), the number of frequent collocates is low. The KWICGrouper gathers the lines with the most hits at the top of the concordance so that in Figure 17.1 the two lines with *Lord* and *Chancellor* are moved up. In the collocation plot (Figure 17.1, top) the semantic connections between *fog* and its collocates remain abstract, while the KWICGrouped concordance lines (Figure 17.1, bottom) provide the collocations in context and point to further textual similarity, emphasising the

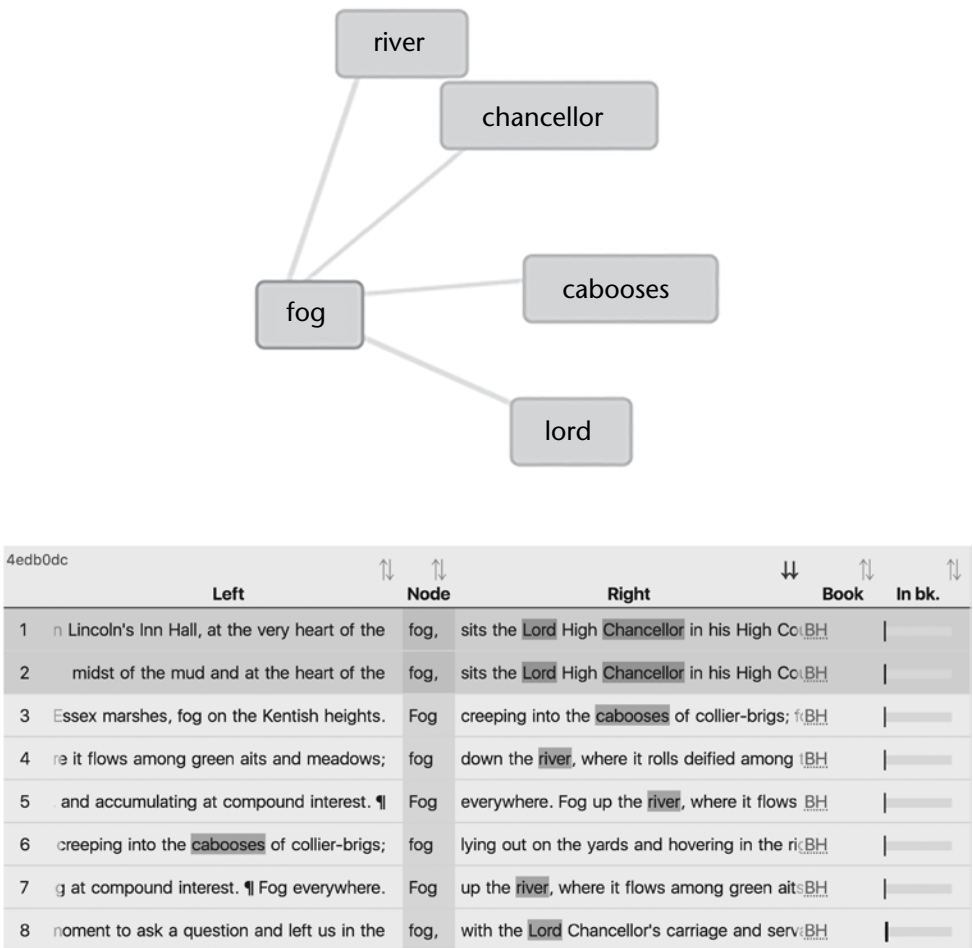


Figure 17.1 Fog collocates in *Bleak House* retrieved with the 'Links' function of Voyant Tools (top) and their context illustrated with the CLiC KWICGrouper (bottom), both within a context of five words around fog

extent of the verbatim repetition in Chapter 1 of *Bleak House*. Examples (2) and (3) provide the extended context of lines 1 and 2 of Figure 17.1, which appear several paragraphs apart in the chapter. Hence, KWICGrouping is a way of visualising textual patterns – somewhere between the raw concordance lines and the more abstract form of collocational networks.

- (2) And hard by Temple Bar, in Lincoln's Inn Hall, at the very heart of the **fog**, sits the Lord High Chancellor in his High Court of Chancery.
(*Bleak House*, Charles Dickens, Chapter 1)
- (3) Thus, in the midst of the mud and at the heart of the **fog**, sits the Lord High Chancellor in his High Court of Chancery.
(*Bleak House*, Charles Dickens, Chapter 1)

Figure 17.2 WordWanderer collocate visualisation of *fog* in *Bleak House* overall (top panel) and in the comparison view with *lord* (bottom panel)

In addition to informing literary investigations of texts, digital tools can provide an entry point to engage with texts in different ways. Pedagogic motivations have played a key role in the development of the CLiC web app (for proposals for its use in the classroom see Mahlberg and Stockwell 2016; Mahlberg et al. 2017). Another example of a tool that allows users to engage with texts is the WordWanderer app. It was created to enable a ‘playful’ navigation through the lexical links in a text and is aimed particularly at non-expert users (Dörk and Knight 2015: 84). The top panel of Figure 17.2 shows the WordWanderer word cloud view of *fog* in *Bleak House* among its collocates. The interactive web app allows

the user to click on the words and their collocates to ‘wander’ through the text in a word cloud supplemented with concordance lines. When selected in the cloud, the collocate is highlighted in the concordance lines, similarly to the CLiC KWICGrouper, although lines are not regrouped. By dragging a word onto another one to create a linking line, the user activates a comparison view, as for *fog* and *lord* in the bottom panel of Figure 17.2. Accordingly, *November*, for instance, is more closely associated with *fog*, whereas *chancellor*, as seen in the examples earlier, is more closely associated with *lord*.

Conventional patterns in narrative

As the example of *fog* in *Bleak House* has illustrated, the textual location of words and patterns plays an important role for the creation of meanings in texts. This is not only true for literary texts. O’Donnell et al. (2012) study newspaper articles and develop a methodology to identify patterns that tend to occur in text-initial positions, such as introductory sentence patterns around the adjective *fresh* that highlight the newsworthiness of a story (. . . *faced fresh embarrassment over* . . .). The opening of a story is particularly crucial when a fictional world is created. Hoey (1983: 6–9) uses the example *once upon a time* to illustrate the power of lexical signals for structuring narratives. He carries out an experiment related to the parlour game ‘Consequences’ with the modification that each participant gets to see one previous sentence before producing the next sentence. The game is started with a prompt sentence. Hoey (1983) runs the experiment several times with different prompts (and different groups), from four sources: two essays by different writers, a short story and a fairy tale. It turns out that only the two games prompted by the first sentence from the fairy tale, *The Flying Trunk*, produce meaningful stretches of discourse: ‘Once upon a time there was a merchant; a man so rich that he could have paved the whole street and most of a small alley with silver’ (Hoey 1983: 8). The formulaic first sentence makes the fairy tale easier to recognise (also see Toolan 2016: Chapter 2).

The way in which lexical signals or conventional phrases function in narrative becomes particularly apparent when their uses are compared across a set of texts. Concordance 17.2 lists occurrences of *once upon a time* in a sample of children’s books. In lines 10, 11, 13 and 17 the phrase is found at the very beginning of the book. This is visible in the last column, where the slider indicates the position in the text. The text preceding the sentences with *once upon a time* in these lines comes from the book or chapter titles. Although line 17 looks different, the text before the initial sentence is a Wordsworth quotation, before the story of the *Water-babies* begins with ‘Once upon a time there was a little chimney-sweep, and his name was Tom’. These are examples of how the meaning of *once upon a time* is associated with its position in the text, which seems to be especially the case for stories written for children.

Example (4) illustrates another meaning of *once upon a time* – not as a story-opening signal, but as a more general time adverbial. This is a usage that present-day readers will be less familiar with:

- (4) “No, no; it’s not so bad as that, my boy. I’ve better eyes than most people, and then I had the privilege of knowing your excellent father rather well *once upon a time*. You haven’t studied his little peculiarities closely enough; but you’ll improve. By the way, where is your excellent father all this time?”

(*Vice Versa*, F. Anstey, Chapter 18)

8	his younger brother, which is every word of it true. ¶	Once upon a time,	certainly as long ago as the first man, or perhaps	howwhy	
9	him the same things. Seals had come to those islands	once upon a time,	but men had killed them all off. Even when he	jungle	
10	The Tale of Two Bad Mice Beatrix Potter ¶ BOOK 1. ¶	ONCE upon a time	there was a very beautiful doll's-house; it was red brick	mice	
11	CHAPTER I. How the Fairies were not invited to Court. ¶	ONCE upon a time	there reigned in Pantouflla a king and a queen. With	priglo	
12	the bank, he told the story of The Devoted Friend. ¶	"Once upon a time,"	said the Linnet, "there was an honest little fellow named	prince	
13	The Tale of Peter Rabbit Beatrix Potter ¶ BOOK 1. ¶	Once upon a time	there were four little Rabbits, and their names were-- ¶ Flop: rabbit		
14	and crossbones. Hugh John knew this ensign in a moment.	Once upon a time,	in his wild youth, he had served under it as	toadydion	
15	blue jacket of a dark shade in which he had	once upon a time	rowed a famous race. It hung now in a glass	toadydion	
16	ad the privilege of knowing your excellent father rather well	once upon a time.	You haven't studied his little peculiarities closely enough; bu: vice		
17	to think, What man has made of man." ¶ WORDSWORTH. ¶	Once upon a time	there was a little chimney-sweep, and his name was Tom	water	

Concordance 17.2 Sample of instances of *once upon a time* in the 19th Century Children's Literature Corpus (ChLit; 10 out of 21 examples)

Concordances for *once upon a time* in the CLiC 19C corpus, a more general corpus than the children's literature corpus, do not include text-opening examples. The phrase is used either as a general time adverbial, or when a story is told at some point in the course of the narrative.

Dickens's *A Christmas Carol* contains a specific example of *once upon a time* beginning a story, but not the text. The phrase appears a couple of pages into the narrative. The narrator starts by talking about the fact that Marley is dead and by describing Scrooge as a miserly and unfeeling person. Both points are crucial preparation for the ghostly story where Scrooge will see Marley's face in the door knocker and eventually will go through a complete transformation of his character. This story starts with *once upon a time* (Example 5) and continues through to the final pages where the narrator concludes the book.

- (5) *Once upon a time*--of all the good days in the year, on Christmas Eve--old Scrooge sat busy in his counting-house. It was cold, bleak, biting weather: foggy withal: and he could hear the people in the court outside, go wheezing up and down, beating their hands upon their breasts, and stamping their feet upon the pavement stones to warm them.

(*A Christmas Carol*, Charles Dickens, Stave I)

Specific and general language use

The previous two sections have focused on how corpus methods can support the analysis of specific qualities of literary texts. The first section has demonstrated how the repeated occurrence of words creates meanings and patterns in a specific fictional world, while the second section looked at conventional fictional patterns. As Carter (2004) has argued, the distinction between literary and non-literary language is not a hard and fast one, and 'literariness' is best seen along a cline. Corpus linguistic methods can also help to contextualise features of literary texts so that it is possible to see whether they are rather idiosyncratic or relate to more general language use. An example is the speech of fictional characters. Specific phrases can be used to individualise characters, and Charles Dickens is an author who is often seen as particularly skillful in this regard. Mahlberg (2013) employed the method of generating clusters, that is, repeated sequences of words such as *what do you mean by*, to identify features of character speech. The sequence *put the case that* is an example of a four-word cluster. It is exclusive to Mr. Jaggers in *Great Expectations* and specifically occurs in Chapter 51. This cluster does not occur in any of the other CLiC corpora either. Even in the

4ed50dc	Left	Node	Right	Book	In bk.
1	nothing. ¶ "Now, Pip," said Mr. Jaggers, "put this case.	Put the case that	a woman, under such circumstances as you have menti	GE	
2	of his defence, how the fact stood about that child.	Put the case that	at the same time he held a trust to find	GE	
3	lady to adopt and bring up." ¶ "I follow you, sir." ¶	"Put the case that	he lived in an atmosphere of evil, and that all	GE	
4	ing generated in great numbers for certain destruction.	Put the case that	he often saw children solemnly tried at a criminal bar	GE	
5	criminal bar, where they were held up to be seen;	put the case that	he habitually knew of their being imprisoned, whipped,	GE	
6	ways for the hangman, and growing up to be hanged.	Put the case that	pretty nigh all the children he saw in his daily	GE	
7	too; if you are lost, your child is still saved."	Put the case that	this was done, and that the woman was cleared." ¶ "I	GE	
8	of the world and went to him to be sheltered.	Put the case that	he took her in, and that he kept down the	GE	
9	. Do you comprehend the imaginary case?" ¶ "Quite." ¶	"Put the case that	the child grew up, and was married for money. That	GE	

Concordance 17.3 All instances of *put the case that* in the CLiC corpora (all appear in *Great Expectations*)

BNC, *put the case that* only occurs once, in a sample from the Six O’Clock News. Whereas the broadcast uses the cluster to recount a past event – see Example (6) – Jaggars’s repeated use of the cluster illustrates his train of thought (cf. Concordance 17.3). (All BNC data were accessed via the BNCweb at <http://bncweb.lancs.ac.uk/>.)

- (6) Michael Heseltine let his deputy Tim Eggar *put the case that* the government had tried to encourage the market for domestic coal.

(BNC, file K6F)

Example (7) provides the context for concordance lines 1–6.

- (7) “Now, Pip,” said Mr. Jaggers, “put this case. *Put the case that* a woman, under such circumstances as you have mentioned, held her child concealed, and was obliged to communicate the fact to her legal adviser, on his representing to her that he must know, with an eye to the latitude of his defence, how the fact stood about that child. *Put the case that* at the same time he held a trust to find a child for an eccentric rich lady to adopt and bring up.”

"I follow you, sir."

“Put the case that he lived in an atmosphere of evil, and that all he saw of children, was, their being generated in great numbers for certain destruction. *Put the case that* he often saw children solemnly tried at a criminal bar, where they were held up to be seen; *put the case that* he habitually knew of their being imprisoned, whipped, transported, neglected, cast out, qualified in all ways for the hangman, and growing up to be hanged. *Put the case that* pretty nigh all the children he saw in his daily business life, he had reason to look upon as so much spawn, to develop into the fish that were to come to his net – to be prosecuted, defended, forsworn, made orphans, bedevilled somehow.”

(*Great Expectations*, Charles Dickens, Chapter 51)

As a cluster in a very specific situation that emphasises Jaggars's role as a member of the legal profession, *put the case that* can be seen at one end of the cline of literariness, where phrases are used to individualise fictional characters. Using corpus methods, we can also identify forms at the opposite end of this cline that are more common in both fictional speech and general spoken language (see Mahlberg et al. 2019). The cluster *it seems to me*

that is such an example – Concordances 17.4 and 17.5 show that this cluster appears both in the BNC Spoken and in 19C to mark personal opinion but with a tendency to hedge or function as vague language. Note, for example, the hesitation markers *er* and *erm* co-occurring with the cluster in the BNC (Concordance 17.4, lines 4 and 6). As a hedge, the cluster can be used to soften negative statements. Example (8) provides the extended context of Concordance 17.4, line 1, where *it seems to me that* indicates potential criticism and ‘a need to review plans’.

- (8) Now it may well be that the whole balance of things is that we’re generally okay for the moment but *it seems to me that* as time goes on and there’s a need to review plans and there’s a need to make further provision that it would be very very difficult indeed against a blanket policy as opposed to individual justifications around ind--er individual settlements.

(BNC, file J9V)

- (9) “*It seems to me that* her nose isn’t quite straight,” said Johnny Eames. Now, it undoubtedly was the fact that the nose on Mrs Lupe’s face was a little awry.

(*The Small House at Allington*, Anthony Trollope, Chapter 4)

A fictional example of the cluster softening an observation is given in (9), which extends the context of line 6 in Concordance 17.5, from *The Small House at Allington*. In line 2 of Concordance 17.5, the cluster appears with even harsher criticism that is indicated in the context by the reporting clause *asked Sir Henry sharply* (from Conan Doyle’s *The Hound of the Baskervilles*).

When comparing authentic spoken language from the twentieth century, as recorded in the BNC, with fictional speech from the nineteenth century, we are not looking for a perfect match. There is no general corpus of nineteenth-century speech comparable to the BNC. Instead, we take the BNC as a ‘proxy for linguistic background knowledge of today’s reader’ (Mahlberg and Wiegand 2018: 131). Accordingly, the textual patterns that are common to fictional speech and authentic spoken language help us to model a possible ‘effect of spokenness in the reader’ (Mahlberg and Wiegand 2018: 132).

The three types of examples that we have shown highlight that corpus methods can help to trace and compare linguistic features across texts and corpora, but also facilitate close

1	okay for the moment but	it seems to me that	as time goes on and
2	and Iris Murdoch, and	it seems to me that	at least part of their
3	all of us today,	it seems to me that	the, that that which
4	Madam. Er er and	it seems to me that	the court should n’t be
5	who hated a woman.	It seems to me that	the discussion right now is
6	U D Ps. Erm	it seems to me that	the emphasis has changed slightly
7	all the factors. So	it seems to me that	these were a set of
8	and in the court.	It seems to me that	we have not a very
9	further use to us.	It seems to me that	’s an analogy with natural
10	Freud insisted on it.	It seems to me that	, what, what ’s

Concordance 17.4 Random sample of *it seems to me that* in the BNC Spoken (from a total of 161 hits in 77 different texts)

4ed0b0dc	Left	Node	Right	Book	In bk.
1	at, if any other punishment had taken its place. But	it seems to me that	a blackguard such as Crosbie can escape now altogetalli		
2	atural." ¶ "What business?" asked Sir Henry sharply.	"It seems to me that	all you gentlemen know a great deal more than I	basker	
3	, Mr Gresham, ¶ "I wonder how Dumbello likes it." ¶	"It seems to me that	Dumbello takes it very easily." ¶ "There are some m	alli	
4	try?" ¶ "Well then, Sir, I accounts for it this way.	It seems to me that	'ere wolf escaped--simply because he wanted to ge	dracula	
5	sensations may die away, just when we want it most.	It seems to me that	her imagination is beginning to work. Whilst she has	dracula	
6	Mr Cradell did not understand much about form. ¶	"It seems to me that	her nose isn't quite straight," said Johnny Eames. N	alli	
7	conceit. Of course I'll go if you ask me, but	it seems to me that	I'm made to knock under to everybody. I hear a	alli	
8	with sudden interest. 'What about it?' ¶ 'Why, sir,	it seems to me that	I have got a pretty distinct chain of evidence, inculp	NorthS	
9	"Being to all intents and purposes a prisoner here,	it seems to me that	I have no choice but to try the old prison	arma	
10	'd, now, gentlemen," said the baronet with decision,	"it seems to me that	I have spoken quite enough about the little that I	basker	

Concordance 17.5 Sample lines of *it seems to me that* in the 19C corpus (10 out of 32 examples)

reading by zooming in on very specific examples. As Simpson (2014: 100) states: ‘[i]t is a curious feature of the corpus method that it simultaneously offers both the micro-detail of a text’s composition and a bird’s eye view of the broader patterns in language that often elude direct observation’.

Future directions

We conclude this chapter by suggesting three main opportunities for literary stylistics: (1) bringing approaches in corpus linguistics and digital humanities more closely together; (2) contributing to developing and critically reviewing visualisation techniques; and (3) applying the outcomes of (1) and (2) to pedagogic practice in the digital world. As we have outlined in this chapter, literary stylistics has seen a corpus turn. One of the opportunities that corpus linguistics brings to literary stylistics is to relate patterns in literary texts to wider patterns of general language use so that the relationships between fictional worlds and the real world can be better understood. At the same time, quantitative approaches in the digital humanities provide new insights into the cumulative picture that is created by looking at lots of literary texts so that the relationships between lots of fictional worlds can be better understood. An important area of future work is to relate the detailed focus on individual texts that literary stylistic analysis can provide with both the big picture views provided by corpus linguistics and digital humanities. The transparent explanations that characterise literary stylistic analysis, a point Toolan (2014) highlights in his list (cf. point 7), will help in this process and publications like the present hand-book are an important contribution to adding transparency to our scientific discourses more widely.

Reading, and how visualisation helps to read differently, is central to how texts are approached both in the digital humanities and in corpus linguistics. Literary stylistics can make an important contribution by guiding new methodological developments in this area. In this chapter, we briefly touched upon the relationship between patterns in concordances and related patterns in network visualisations. A better understanding of this relationship will be crucial for visualisation tools that can support meaningful text analysis. Insights from literary stylistics into the importance of links between meanings and the places where meanings are created in texts (opening chapters, the speech of individual characters, etc.) will also be vital in this regard.

Visualisation does not only affect the way texts and data are analysed, but also how research is shared. An example is the BRANCH project (www.branchcollective.org/). This website publishes peer-reviewed articles about events in the nineteenth century that are interactively plotted along a visual timeline. BRANCH also encourages contributors to critically reflect on issues of the linearity of timelines, and, as the editor-in-chief explains, the project aims ‘to rethink our approach to scholarly publication itself through the tools made available by the digital humanities’ (Felluga 2013). This website relates to the COVE initiative (<http://covecollective.org>), which hosts open-access digital tools for Victorian studies and annotated, scholarly editions of literary works (Felluga 2015). Such initiatives can complement the range of corpora and software that corpus stylisticians have been drawing on so far. They also highlight that visualisation is not only relevant to advance scholarly knowledge of linguistic features in literature, but, importantly, has great potential in the dissemination of such knowledge.

The way in which literary stylisticians embrace the opportunities of bringing disciplines together and developing new methodologies will have direct implications for pedagogy. As we have argued, in the digital humanities, ‘reading’ is often reconceptualised in the light of new methodological developments. The practice of reading itself provides a link to pedagogy. Picton and Clark (2015), for instance, report on a study that shows how e-books affect reading motivation and skills of children. With the increasing variety of formats in which literature can be accessed and the ubiquity of digital activity more widely, there is a major opportunity for literary stylistics to consider how new methods of analysis can also account for and support new forms of reading. In this chapter, we specifically used the CLiC web app to illustrate how digital approaches provide methods for the analysis of literary texts. Such methods can also contribute new dimensions to the reading experience more widely. As Zyngier and Fialho (2015: 219) point out, literary texts are ‘an optimal locus for experiencing humanity’. Hence in learning environments, literary stylistics has the opportunity to bring together the experiences of learners in the digital world with new approaches to how they analyse, experience and contextualise literary texts. Tools like WordWanderer and CLiC are designed to be used beyond academic research projects and make their way into secondary classrooms. To facilitate this, there is a need for introductory resources. The freely accessible *CLiC Activity Book* (Mahlberg et al. 2017) is an example of such a resource that provides teachers with a repertoire of corpus stylistic classroom activities.

Further reading

- 1 McIntyre, D. and Walker, B. (2019). *Corpus stylistics*. Edinburgh: Edinburgh University Press.
This recent textbook provides an overview of corpus stylistics, covering theoretical backgrounds, points of debate and practical instructions. It includes analyses of both literary and non-literary language.
- 2 Leech, G. and Short, M. ([1981] 2007). *Style in fiction: A linguistic introduction to English fictional prose*, 2nd ed. Harlow: Pearson.
This second edition of Leech and Short’s seminal book contains a chapter reviewing significant developments in the discipline of stylistics since the first publication of this book.
- 3 Simpson, P. ([2004] 2014). *Stylistics: A resource book for students*, 2nd ed. London: Routledge.
In this second edition, Paul Simpson has added sections on recent trends in the field, including a discussion of and case study from corpus stylistics. As a resource book, it is designed so that students can start with the foundations and build up their knowledge of specific areas by reading the ‘extending’ sections which contain reprints of published stylistic research by a variety of scholars.

References

- Anthony, L. (2018). Visualisation in corpus-based discourse studies. In C. Taylor and A. Marchi (eds.), *Corpus approaches to discourse: A critical review*. Abingdon: Routledge, pp. 197–224.
- Balossi, G. (2014). *A corpus linguistic approach to literary language and characterization: Virginia Woolf's The Waves*. Amsterdam: John Benjamins.
- Best, S. and Marcus, S. (2009). Surface reading: An introduction. *Representations* 108(1): 1–21. doi:10.1525/rep.2009.108.1.1.
- Brezina, V., McEnery, T. and Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20(2): 139–173. doi:10.1075/ijcl.20.2.01bre.
- Burke, M. (ed.). (2014). *The Routledge handbook of stylistics*. London: Routledge.
- Burnard, L. (2014). *What is the text encoding initiative? How to add intelligent markup to digital resources*. Marseille: OpenEdition Press. Available at: <http://books.openedition.org/oep/426> (Accessed April 2019).
- Burrows, J. (2002). “Delta” – A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17(3): 267–287.
- Busse, B. (2010). *Speech, writing and thought presentation in a corpus of 19th-century narrative fiction*. Bern: University of Bern.
- Carter, R. (2004). *Language and creativity: The art of common talk*. London: Routledge.
- Carter, R. (2010). Issues in pedagogical stylistics: A coda. *Language and Literature* 19(1): 115–122. doi:10.1177/0963947009356715.
- Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*. *International Journal of Corpus Linguistics* 14(1): 29–59. doi:10.1075/ijcl.14.1.03cul.
- Dörk, M. and Knight, D. (2015). WordWanderer: A navigational approach to text visualization. *Corpora* 10(1): 83–94.
- Eder, M. (2017). Visualization in stylometry: Cluster analysis using networks. *Digital Scholarship in the Humanities* 32(1): 50–64. doi:10.1093/lle/fqv061.
- Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: A package for computational text analysis. *The R Journal* 8(1): 107–121.
- Emmott, C. and Alexander, M. (2014). Foregrounding, burying and plot construction. In P. Stockwell and S. Whiteley (eds.), *The Cambridge handbook of stylistics*. Cambridge: Cambridge University Press, pp. 329–343.
- Felluga, D.F. (2013). BRANCHing out: Victorian studies and the digital humanities. *Critical Quarterly* 55(1): 43–56. doi:10.1111/crq.12032.
- Felluga, D.F. (2015). The eventuality of the digital. *19: Interdisciplinary Studies in the Long Nineteenth Century* 2015(21). Available at: <https://19.bbk.ac.uk/article/id/1693/> (accessed April 2019).
- Galleron, I. (2017). Conceptualisation of theatrical characters in the digital paradigm: Needs, problems and foreseen solutions. *Human and Social Studies* 6(1): 88–108. doi:10.1515/hssr-2017-0007.
- Galleron, I., Idmhand, F., Goloubkoff, A., Meynard, C., Roger, J. and Wildocher, A. (2018). Spotting the character: How to collect elements of characterisation in literary texts? In *DH2018*, Mexico City. Available at: <https://dh2018.adho.org/en/spotting-the-character-how-to-collect-elements-of-characterisation-in-literary-texts/> (Accessed April 2019).
- Gavins, J. (2007). *Text world theory: An introduction*. Edinburgh: Edinburgh University Press.
- Gavins, J. and Steen, G. (2003). *Cognitive poetics in practice*. London: Routledge.
- Giovanelli, M. (2018). “Something happened, something bad”: Blackouts, uncertainties and event construal in *The Girl on the Train*. *Language and Literature* 27(1): 38–51. doi:10.1177/0963947017752807.
- Hall, G. (2014). Stylistics as literary criticism. In P. Stockwell and S. Whiteley (eds.), *The Cambridge handbook of stylistics*. Cambridge: Cambridge University Press, pp. 87–100.

- Ho, Y-F., Lugea, J., McIntyre, D., Xu, Z. and Wang, J. (2018). Text-world annotation and visualization for crime narrative reconstruction. *Digital Scholarship in the Humanities*. doi:10.1093/llc/fqy044
- Hockey, S. (2000). *Electronic texts in the humanities: Principles and practice*. Oxford: Oxford University Press.
- Hockey, S. (2004). The history of humanities computing. In S. Schreibman, R. Siemens and J. Unsworth (eds.), *A companion to digital humanities*. Oxford: Blackwell, pp. 3–19. Available at: www.digitalhumanities.org/companion/ (Accessed April 2019).
- Hoey, M. (1983). *On the surface of discourse*. London: George Allen & Unwin.
- Hoover, D.L. (2007). Corpus stylistics, stylometry, and the styles of Henry James. *Style* 41(2): 174–203.
- Jeffries, L. (1996). What makes language into art? In J. Maybin and N. Mercer (eds.), *Using English: From conversation to canon*. London: Routledge, pp. 162–184.
- Jeffries, L. and Walker, B. (2018). *Keywords in the press: The New Labour years*. London: Bloomsbury.
- Jensen, K.E. (2014). Linguistics and the digital humanities: (Computational) corpus linguistics. *MedieKultur* 57: 115–134.
- Jockers, M.L. (2013). *Macroanalysis: Digital methods and literary history*. Champaign: University of Illinois Press.
- Kirschenbaum, M.G. (2010). What is digital humanities and what's it doing in English departments? *ADE Bulletin* 150: 55–61. doi:10.1632/ade.150.55.
- Klein, L. (2018). *Distant reading after Moretti*. Available at: <https://lklein.com/2018/01/distant-reading-after-moretti/> (Accessed April 2019).
- Lee, J.J., Greteman, B., Lee, J. and Eichmann, D. (2018). Linked reading: Digital historicism and early modern discourses of race around Shakespeare's Othello. *Journal of Cultural Analytics*. doi:10.22148/16.018.
- Leech, G. (1985). Stylistics. In T.A. van Dijk (ed.), *Discourse and literature: New approaches to the analysis of literary genres*. Amsterdam: John Benjamins, pp. 39–57.
- Leech, G. (2008). *Language in literature: Style and foregrounding*. Harlow: Pearson Education.
- Leech, G. and Short, M. ([1981] 2007). *Style in fiction: A linguistic introduction to English fictional prose*, 2nd ed. Harlow: Pearson.
- Mahlberg, M. (2013). *Corpus stylistics and Dickens's fiction*. London: Routledge.
- Mahlberg, M., Conklin, K. and Bisson, M.-J. (2014). Reading Dickens's characters: Employing psycholinguistic methods to investigate the cognitive reality of patterns in texts. *Language and Literature* 23(4): 369–388.
- Mahlberg, M. and McIntyre, D. (2011). A case for corpus stylistics: Ian Fleming's *Casino Royale*. *English Text Construction* 4(2): 204–227.
- Mahlberg, M. and Stockwell, P. (2016). Point and CLiC: Teaching literature with corpus stylistic tools. In M. Burke, O. Fialho and S. Zyngier (eds.), *Scientific approaches to literature in learning environments*. Amsterdam: John Benjamins, pp. 251–267.
- Mahlberg, M., Stockwell, P., Joode, J. de, Smith, C. and O'Donnell, M.B. (2016). CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics. *Corpora* 11(3): 433–463. doi:10.3366/cor.2016.0102.
- Mahlberg, M., Stockwell, P. and Wiegand, V. (2017). *CLiC – corpus linguistics in context: An activity book. Supporting the teaching of literature at GCSE and AS/A-level*. Available at: <https://birmingham.ac.uk/clic-activity-book> (Accessed April 2019).
- Mahlberg, M. and Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse and N. Nørgaard (eds.), *Rethinking language, text and context: Interdisciplinary research in stylistics in honour of Michael Toolan*. London: Routledge, pp. 123–143.
- Mahlberg, M., Wiegand, V., Stockwell, P. and Hennessey, A. (2019). Speech-bundles in the 19th-century English novel. *Language and Literature* 28(4): 326–353. doi: 10.1177/0963947019886754.
- Mee, J. (2010). *The Cambridge introduction to Charles Dickens*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511780646.

- Meister, J.C., Horstmann, J., Petris, M., Jacke, J., Bruck, C., Schuhmacher, M. and Flüh, M. (2019). *CATMA 6.0.0*. doi:10.5281/zenodo.3523228.
- Mendenhall, T.C. (1901). A mechanical solution of a literary problem. *Popular Science Monthly*: 60. Available at: https://en.wikisource.org/wiki/Popular_Science_Monthly/Volume_60/December_1901/A_Mechanical_Solution_of_a_Literary_Problem (Accessed January 2019).
- Michel, J-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A. and Aiden, E.L. (2011). Quantitative analysis of culture using millions of digitized books. *Science* 331(6014): 176–182. doi:10.1126/science.1199644.
- Moretti, F. (2013). *Distant reading*. London: Verso.
- Nørgaard, N., Busse, B. and Montoro, R. (2010). *Key terms in stylistics*. London: Bloomsbury.
- Nyhan, J., Terras, M. and Vanhoutte, E. (2013). Introduction. In M. Terras, J. Nyhan and E. Vanhoutte (eds.), *Defining digital humanities: A reader*. Farnham: Ashgate, pp. 1–10.
- O'Donnell, M.B. (2008). KWICgrouper – designing a tool for corpus-driven concordance analysis. *International Journal of English Studies* 8(1): 107–122.
- O'Donnell, M.B., Scott, M., Mahlberg, M. and Hoey, M. (2012). Exploring text-initial words, clusters and congrams in a newspaper corpus. *Corpus Linguistics and Linguistic Theory* 8(1): 73–101.
- Pechenick, E.A., Danforth, C.M. and Dodds, P.S. (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS One* 10(10): e0137041. doi:10.1371/journal.pone.0137041.
- Picton, I. and Clark, C. (2015). *The impact of ebooks on the reading motivation and reading skills of children and young people: A study of schools using RM books*. London: National Literacy Trust. Available at: <https://literacytrust.org.uk/research-services/research-reports/impact-ebooks-reading-motivation-and-reading-skills-children-and-young-people/> (Accessed April 2019).
- Pope, C.W. (2010). The bootcamp discourse and beyond. *International Journal of Corpus Linguistics* 15(3): 323–325. doi:10.1075/ijcl.15.3.01wor.
- Schwarzbach, F.S. (1990). *Bleak House: The social pathology of urban life*. *Literature and Medicine* 9(1): 93–104. doi:10.1353/lm.2011.0162.
- Semino, E. and Culpeper, J. (2002). *Cognitive stylistics: Language and cognition in text analysis*. Amsterdam: John Benjamins.
- Semino, E. and Short, M. (2004). *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing*. London: Routledge.
- Short, M. (1996). *Exploring the language of poems, plays, and prose*. London: Longman.
- Simpson, P. ([2004] 2014). *Stylistics: A resource book for students*, 2nd ed. London: Routledge.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sinclair, J. (2003). *Reading concordances: An introduction*. Harlow: Longman.
- Sinclair, S. and Rockwell, G. (2015). Text analysis and visualization: Making meaning count. In S. Schreibman, R. Siemens and J. Unsworth (eds.), *A new companion to digital humanities*. Chichester: John Wiley & Sons, pp. 274–290. doi:10.1002/9781118680605.ch19.
- Sinclair, S. and Rockwell, G. (2018). *Voyant Tools*. Available at: <http://voyant-tools.org> (Accessed April 2019).
- Sinclair, S. and Rockwell, G. (n.d.). *Principles of Voyant Tools*. Available at: <http://docs.voyant-tools.org/context/principles/> (Accessed April 2019).
- Sorlin, S. (2016). Three major handbooks in three years: Stylistics as a mature discipline. *Language and Literature* 25(3): 286–301. doi:10.1177/0963947016660221.
- Sotirova, V. (ed.). (2015). *The Bloomsbury companion to stylistics*. London: Bloomsbury Academic.
- Spitzer, L. (1948). *Linguistics and literary history: Essays in stylistics*. Princeton: Princeton University Press.
- Stockwell, P. (2002). *Cognitive poetics: An introduction*. London: Routledge.
- Stockwell, P. (2006). Language and literature: Stylistics. In B. Aarts and A. McMahon (eds.), *The handbook of English linguistics*. Oxford: Blackwell, pp. 742–758. doi:10.1002/9780470753002.ch31.
- Stockwell, P. (2010). The eleventh checklist of the apocalypse. In B. Busse and D. McIntyre (eds.), *Language and style: In honour of Mick Short*. Basingstoke: Palgrave Macmillan, pp. 419–432.

- Stockwell, P. and Whiteley, S. (2014a). Introduction. In P. Stockwell and S. Whiteley (eds.), *The Cambridge handbook of stylistics*. Cambridge: Cambridge University Press, pp. 1–9.
- Stockwell, P. and Whiteley, S. (eds.). (2014b). *The Cambridge handbook of stylistics*. Cambridge: Cambridge University Press.
- Stockwell, P. and Mahlberg, M. (2015). Mind-modelling with corpus stylistics in *David Copperfield*. *Language and Literature* 24(2): 129–147. doi:10.1177/0963947015576168.
- Terras, M. (2010). Present, not voting: Digital humanities in the panopticon. In *Digital humanities 2010*. London: King's College. Available at: <http://melissaterras.blogspot.com/2010/07/dh2010-plenary-present-not-voting.html> (Accessed April 2019).
- Terras, M. (2016). A decade in digital humanities. *Journal of Siberian Federal University: Humanities & Social Sciences* 9(7): 1637–1650.
- Terras, M., Gray, S. and Ammann, R. (2013). Textal: A text analysis smartphone app for digital humanities [Poster]. In *Digital humanities 2013*. Lincoln: University of Nebraska. Available at: <http://dh2013.unl.edu/abstracts/ab-227.html> (Accessed April 2019).
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.
- Toolan, M. (2014). The theory and philosophy of stylistics. In P. Stockwell and S. Whiteley (eds.), *The Cambridge handbook of stylistics*. Cambridge: Cambridge University Press, pp. 13–31.
- Toolan, M. (2016). *Making sense of narrative text: Situation, repetition, and picturing in the reading of short stories*. New York: Routledge.
- Underwood, T. (2017). A genealogy of distant reading. *Digital Humanities Quarterly* 11(2). Available at: www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html (Accessed April 2019).
- Underwood, T., Bamman, D. and Lee, S. (2018). The transformation of gender in English-language fiction. *Journal of Cultural Analytics*. doi:10.31235/osf.io/fr9bk.
- Wales, K. (2001). *A dictionary of stylistics*. Harlow: Pearson Education.
- Wales, K. (2014). The stylistic tool-kit: Methods and sub-disciplines. In P. Stockwell and S. Whiteley (eds.), *The Cambridge handbook of stylistics*. Cambridge: Cambridge University Press, pp. 32–45.
- Wang, J., Ho, Y., Xu, Z., McIntyre, D. and Lugea, J. (2016). The visualisation of cognitive structures in forensic statements. In *Proceedings – 2016 20th International Conference Information Visualisation, IV 2016*. Lisbon, pp. 106–111. doi:10.1109/IV.2016.60.
- Werth, P. (1999). *Text worlds: Representing conceptual space in discourse*. Harlow: Longman.
- Whiteley, S. and Canning, P. (2017). Reader response research in stylistics. *Language and Literature* 26(2): 71–87. doi:10.1177/0963947017704724.
- Zyngier, S. and Fialho, O. (2015). Pedagogical stylistics: Charting outcomes. In V. Sotirova (ed.), *The Bloomsbury companion to stylistics*. London: Bloomsbury Academic, pp. 208–230.

Historical linguistics

Freek Van de Velde and Peter Petr 

Introduction

The bulk of this chapter is devoted to two major issues. First, we deal with the types of data (corpora) that have been mined with the assistance of computers in historical linguistic research and to the criteria related to good corpus design, including sampling practices, representativity issues, dating, the impact of text genres and spelling variation. Much of these requirements are fairly well-known and also hold for linguistic research on contemporary language and will be outlined only briefly here. Some other aspects are more specific to historical research and deserve to be mentioned separately. We will also discuss the challenges and possibilities of the currently available datasets for historical linguists. Second, we introduce a number of commonly used statistical techniques used for the analysis of language change.

Background

Historical linguistics has, of necessity, always worked with textual material as its empirical basis. Because of this early commitment to attested data, historical linguists were among the pioneers in embracing digital corpora and the ensuing methodologies (see e.g. Hockey 2004), first with concordances on classical works for philological purposes, and later with the systematic retrieval of a sample of data from a principled selection of texts. More recently the rise of digital humanities research has promoted a number of other aspects of digital research, including computer-assisted advanced statistical analysis, or visualisation techniques of big data. Historical corpus linguistics has generally kept up well with advancements in statistics, but it arguably lags behind somewhat in data visualisation.

Critical issues and topics

The current buzz around the fashionable term *digital humanities* might suggest that this approach is a recent development in the humanities. While there is no denying that the quantitative approach of large electronically available datasets indeed has gained momentum in

recent years, it must be pointed out that corpus-driven research has a venerable tradition in literary studies – think of concordances of the Bible, for instance – as well as in linguistics, as for instance in the nineteenth-century lexicographic mega-projects leading to the *Oxford English Dictionary* for English, the *Woordenboek der Nederlandsche taal* for Dutch or the *Deutsches Wörterbuch* for German. For historical linguistics in morphosyntax, the 1980s witnessed an important upsurge, with the advent of the Helsinki Corpus, but even before the 1980s, the added value of using attested data was often very clear. Some earlier qualitative studies, such as Lightfoot's (1979) analysis of the rise of modal auxiliaries in English, which used invented examples, were falsified by corpus research (e.g. Loureiro-Porto 2010), and some of the best descriptive work has been firmly based on textual evidence (e.g. Visser 1963–1973). The increasing technical possibilities of retrieving data from tagged and parsed corpora also allows for new types of research questions to be asked. Big data make it possible to look at change across the lifespan, regularities between individuals and the relation between individual and communal change and, more generally, at the when and why of gradience and gradualness in language change. Linking transcriptions to the scanned originals will increasingly allow for multimodal analysis as well, associating linguistic features to layout and images.

At the same time, however, a digital humanities approach holds some risks. There is an increasing demand on linguistic scholars to acquaint themselves with technical aspects of computer coding and statistics, and this may arguably come at the expense of 'old-school' philological expertise. The availability of corpora makes it unnecessary to get deep, hands-on experience with palaeography or codicology. Moreover, high precision (and recall) in corpus study decrease the need to read extensively through full-text editions of ancient texts. Older scholars sometimes lament the new generations' relative lack of familiarity with the body of old texts and deplore the focus on specific problems, often only those which can easily be extracted from a corpus. Not all of these complaints come from 'Luddites' unwilling to embrace methodological innovation. We should certainly be wary of overenthusiasm about the way historical linguistics is undertaken today. For all the methodological merits of quantitative bird's-eye approaches, philological scrutiny is much needed (see also Fischer 2007). As in other fields of scientific enquiry, progress can only be made by combining different forms of expertise, not necessarily in one individual, but by division of labour in team research. And linguistic research, especially in the historical domain, can in our experience at least, only be done properly by scholars with extensive training in 'traditional' linguistics.

Main research methods

Corpora

An electronic (historical) corpus is a principled selection of texts, enriched with metadata for each text. While this may be difficult to achieve, especially for more meagrely attested languages stages, corpus design should aim at yielding a representative sample of texts, which corresponds as well as possible to the characteristics of the represented population (a particular language variety; cf. Biber 1993). Only then can generalisations from the corpus be assumed to apply to the population. A corpus may also aim to be balanced, including a wide range of variation in the population. However, these two may be at odds, as is apparent when comparing cross-genre corpora (which are more balanced) to single-genre corpora (which may be more representative of that genre). The development of cross-genre historical corpora targeted at linguistics started early on in the 1980s, notably with the Helsinki

Corpus of Historical English Texts (HC, Kytö and Rissanen 1992). The HC covers the period 730–1710, including both prose and verse in a range of genres. Its purpose was to enable more representative data-based longitudinal study of the English language. Metadata were geared towards detailed sociolinguistic studies (although inevitably sparser for older texts). Because data were scarce, particularly prior to 1400, the compilers introduced ‘prototypical text types’, which related genres by group, allowing for a more sensible diachronic analysis than either without any genre division at all, or with one that is too fine-grained to yield robust results.

Around the millennium various initiatives were taken to expand the HC and linguistically annotate it. This resulted in a series of corpora coordinated by Penn and York universities: the York-parsed corpus of Old English prose (YCOE) and the York Poetry corpus (YPC) (cf. Taylor and Warner et al. 2003), and the Penn-parsed corpora of Middle English (PPCME2, Kroch and Taylor 2000), Early Modern (PPCEME) and Late Modern English (PCMBE), as well as Early Modern letters (PCEEC, Taylor et al. 2006), and, most recently, Middle English verse (PCMEP, Zimmermann 2016). Each of these corpora has been annotated for parts of speech as well as syntactically parsed according to the Penn treebank standard (which is based on LISP, a then-common language for NLP applications) (Taylor and Marcus et al. 2003). Of course, such parsing is still selective, reflecting the choices of the researchers who developed the treebank annotation system (cf. also Broccias 2008: 39–40). Certain syntactic structures, such as the middle construction (*This tent sleeps four*), are not explicitly annotated and remain notoriously difficult to retrieve (automatically or manually). In some other cases, tags conflate different categories. For instance, the early English separable and inseparable prefixes *to-* (as in *to-comen* ‘arrive’ and *to-fallen* ‘fall apart’, respectively) both received the tag of ‘separable prefix’. However, these slight inaccuracies are by far outweighed by the enormous usefulness of these corpora. Note that all annotation they contain resulted from a largely manual effort. The natural next step is automated part-of-speech tagging and syntactic parsing, but the problems involved are many: orthographic variation, dialectal diversity, different corpus designs, for example, with regard to tokenisation (e.g. is *thinkstou* one or two tokens?). There have been recent attempts to use the Penn-parsed corpora to train part-of-speech (POS) taggers for tagging EEBO (e.g. Ecay 2015), but the underlying procedures and the resulting accuracy are poorly documented, and it appears that the challenge still stands.

Thanks to these corpora, our knowledge of various aspects of the history of English syntax has been significantly advanced. The chosen design, however, also comes with a number of issues. Because of the focus on syntactic research, the parsed corpora were (mostly) limited to prose texts only. Verse constitutes nevertheless an important share of pre-modern textual production, even more so for those periods in which data are very scarce to start with. As a result, the period 1250–1350 only contains three texts, two of which are clumsy translations: a psalter from Latin and the religious treatise *Ayenbit of Inwit* from the French *Somme le roi* (Roux 2010). Another design feature which requires caution is that it is common for a corpus file to consist of an entire scholarly edition, with only a single set of metadata associated with such a file. However, scholarly editions sometimes collate or combine fragments from different manuscripts, which may widely vary in date of production or dialect. Also, manuscripts often contain different types of texts, which belong to different genres, registers, have different exemplars and so on. Finally, while the different corpora seamlessly connect chronologically, they display internal dialect and genre shifts that compromise their comparability. Many of these shortcomings are notoriously hard to

overcome for Old and Middle English without creating other shortcomings. A continuously incremented multi-genre corpus that avoids most problems for the modern period, starting in 1600, is *A Representative Corpus of English Historical Registers ARCHER* (Yáñez-Bouza 2011). Recently it has also been enriched with POS tags, and Penn Treebank parsing is on the agenda. The triplet known as BLOB, LOB and FLOB (constituting the British segment of the Brown corpus family, cf. Hinrichs et al. 2007 and Leech 2013), which are also syntactically annotated, are compiled with the same philosophy of maximal comparability in mind. Unlike ARCHER, they maximise the temporal homogeneity of the subperiods by limiting each one to a single year (1931, 1960 and 1990).

A *leitmotiv* shared by all historical corpora so far is that ‘small is beautiful’. Only by limiting the corpus size was it possible to provide rich annotation while still accomplishing reasonable representativity. One obvious disadvantage though is that research on these corpora is limited to the most frequent phenomena (e.g. core auxiliaries, determiners, morphology, phonology or basic syntactic structures). When dealing with less common or even downright rare constructions or lexemes, these corpora often fall short, as is, for instance, discussed in detail for the demise of *being to* by Hundt (2014: 5). In the absence of alternatives, this either leads to quantitative studies based on rather fewer instances than one would like, which forces researchers back to more qualitatively oriented research (e.g. Pertejo 1999: 138; Trousdale 2008). Note that an inherent danger when faced with so little data is that one concludes that the case at issue is rare *per se*, and therefore unconventional. However, even if a specific linguistic phenomenon only occurs a handful of times in (parts of), say, the HC, it may very well have still been fully conventionalised (cf. Petré 2014: 5–7). Size also restricts the potential scope of a study. For instance, while *be going to* is very frequent in present-day English, it was obviously quite rare when first appearing in the seventeenth century. This is reflected in Pertejo (1999), based on the HC, which yields a mere eight instances to work with. Because of these size limitations, Nevalainen et al. (2011) only analyse cases that have already advanced in their grammaticalisation path and whose frequency has already reached a critical mass.

One obvious tactic to overcome these limitations is using bigger datasets, although this often means giving up the cleanliness or rich annotation of the smaller ones. The shift to big, or at least bigger data, has been gathering speed since about 2005. Unlike the carefully crafted corpora of before, big data typically refurbish existing electronic sources to better suit the needs of linguistic research. Some corpora were designed with an aim to open access and for that reason only draw on texts available in the public domain. Well-known examples are CLMETEV, COHA and EEBOCorp. CLMETEV, the extended corpus of Late Modern English (De Smet 2006, period covered: 1710–1920) converts texts from the Gutenberg Project, among other sources, removing extraneous text and coding, and adding a basic Extended Markup Language (XML) header with structured metadata. The *Corpus of Historical American* (COHA; Davies 2010, 2012) covers the period 1810–2009. COHA is a genre-balanced web-crawled corpus from various online public domain archives. The corpus, though, also illustrates some of the problems involved in big data. Several texts are based on automatic optical character recognition (OCR) transcriptions, and many editions on online archives are reprints (or scholarly editions) rather than first editions, for which publication dates (or even the actual text) may differ widely from that of the original text. COHA provides relatively little information on the amount of post-editing that has been carried out to amend this. While these reliability issues are mitigated by the sheer size of the corpus, error rates may increase the rarer the phenomenon one looks at. EEBOCorp

(Petré 2013a), which covers British English between 1473 and 1700, is with 525 million words similar in size to COHA. It is not, however, balanced in genre. Texts are taken from the open-access part of the TCP-transcriptions of *Early English Books Online* (www.textcreationpartnership.org/). The problem of re-editions has partly been tackled by the automatic exclusion of posthumously published material, but no systematic pruning of reprints has been done. A problem that is shared between COHA (cf. e.g. Hilpert 2011: 438) and EEBOCorp is that the earliest decades are based on significantly less data than the later ones.

Another way of overcoming the limitations of smaller general-purpose corpora is to limit the corpus to a specific subarea of the language. This is not merely a practical solution, but also a theoretically motivated one. The research of Biber et al. (e.g. Biber and Conrad 2009) has amply revealed that different genres come with different language and that they may change over time, both in their properties as well as in their presence (e.g. Biber and Gray 2013). Genre-specific corpora ensure that results are more coherent and also open up new research questions. Letter-corpora, such as ICAMET-letters (1386–1688) or PCEEC (1410–1695), are ideally suited for investigating social networks and language change (Raumolin-Brunberg 2017) as well as specific phenomena such as forms of address (Nevala 2007). Science corpora have been built to look into the development and changing functions of scientific vocabulary and grammar (e.g. the Coruña corpus, cf. Moskowich and Crespo García 2007, the Corpus of Early English Medical Writing, cf. Taavitsainen et al. 2005, 2010), as well as the changing language lay people have used to understand medicine through the ages by means of the *H-CODE*-corpus, currently in progress. Other corpora have been designed specifically for dialectological research. Examples are LAEME (1150–1325), which is also morphosyntactically annotated and lemmatised, and HC of Older Scots (1450–1700).

Recent digitalisation projects also include large amounts of data written by individual authors. Two of the most extensive high-quality databases are *Early English Books Online* (EEBO), covering all available print material from the UK between 1470 and 1700, and the Hansard Corpus (Alexander and Davies 2015), a 1.6-billion corpus containing 7.6 million parliamentary speeches between 1803 and 2005. The availability of these sources has enabled studies in the innovative field of linguistic research into syntactic change at the level of the individual (De Smet 2016; Petré 2017; Petré and Van de Velde 2018). Previously, studies including individual variation and change were limited to phonology or highly frequent grammatical items. With these resources, more detailed insight may be gained in the processes of syntactic innovation, the dynamics of the grammar of individuals and to what extent changing behaviour in their grammar is due to social or cognitive reasons, as well as the spread of innovations through the language community. The *Mind-Bending Grammars* project (2015–2020) specifically focuses on how much syntactic change is attested across the lifespan and how changes interact with each other and with the social context. For this purpose, a 90-million-word corpus of prolific authors from five generations with an active writing career between 1623 and 1757, *Early-Modern Multiloquent Authors* (EMMA, Petré et al. 2019), has been compiled.

Finally, apart from corpora, other types of digital resources have also been compiled and put to various uses. Online dictionaries often contain a wealth of material, and while quantitative research on them is problematic in a number of respects (quotations being truncated, duplicated and typically biased towards the literary canon), when used with care, they may be sensibly used for pilot studies or for the collection of examples that would

otherwise be hard to gather, as is discussed in detail for the *Oxford English Dictionary* by Hoffmann (2004). Online taxonomic dictionaries or thesauri provide opportunities for quantitative research on different lexical types (rather than tokens), such as in Alexander and Kay (2018), who carry out a large-scale analysis of metaphors with reference to the supernatural. Their analysis shows that negative (e.g. *impish* to refer to behaviour) rather than positive (*angelic*, *heavenly*) metaphors are clearly predominant, which they explain in terms of a human urge to shirk responsibility for one's own flaws. Other type- or feature-based databases include *eLALME* (Benskin et al. 2013), which contains detailed information on dialectal features of 1000 Middle English texts, as well as its recent sibling *CoNe: A corpus of Narrative etymologies*, a diachronic database of attested spellings and forms of lexemes contained in the *LAEME* corpus. Both databases allow for quantitative dialectology (e.g. Lass and Laing 2016) and diachronic research into the relation between orthography and phonology (Laing and Lass 2009; Stenbrenden 2015).

The development of historical corpora has taken us on a journey from a million-word corpus stretching across a millennium, to a hundred-million-word corpus covering a mere 135 years, and from multi-genre longitudinal research on highly frequent phenomena to more specific types of research questions (examining individual or genre developments). The discussion of these corpora has inevitably been selective. For a more extensive list of historical corpora of English we refer to the extensive database Cord (www.helsinki.fi/varieng/CoRD/index.html), appropriately hosted by the University of Helsinki, the cradle of English historical corpora. The resources that have been discussed were typically compiled with certain specific research questions in mind. In the meantime, most have also been put to use creatively, to answer research questions that were not envisaged by the compilers. We believe that such creative use of resources is an area in which computer-assisted linguistic research may still make much progress. For instance, Petré (2015) demonstrates how historical databases such as Prosopography of Anglo-Saxon England (PAASE) may shed light on competition between passive auxiliaries, or how corpus-based syntactic analysis might shed light on the mind-set of people from the past, a question that is, for instance, also at the core of the *H-CODE*-corpus and project.

Retrieving data and preparing them for analysis: best practices and pitfalls

Formatting

The best-known format for digital texts is probably the TEI (Text Encoding Initiative), developed from 1987 onwards for literary scholarship. Although this has been changing since about 2005, TEI used to be mostly absent in corpora targeted at historical linguistic research. One of the reasons is that the requirements typical of literary or historical research are significantly different from those set by linguistic analysis. The basic unit of study in literary studies is typically the document in its entirety. In contrast, linguists are primarily interested in words and patterns (combinations) of words, or parts of the document that are coherent texts in terms of genre, date, author, etc. For instance, the preparatory work for the EMMA corpus (see earlier for a description) involved the identification of those parts of a document which were by the author of interest, excluding, for example, contents pages, advertisements, etc., as well as further identification of different textual units within an authored document, which belong to different creation dates or genres.

Correct quantitative text analysis also requires a fairly systematic delineation of the concept of a word. Faithful renditions of the source text, however, do not always yield a text with consistent word delineations. Certain XML tags (e.g. those indicating superscript, subscript or decorative initials) cut into words. A good indexing programme will not be misled by these and will restore word units in those cases. Less straightforward are end-of-line word splits. Some encodings provide specific solutions for this (such as repeating the split word as a single word inside a tag), but not all of them do so. More generally, should hyphens at other positions than the end of a line be considered word separators or indices of compounds? The problem is even greater with apostrophes. A string such as *It's* 'it is' would generally still be considered to be two words, but *king's* in *the king's crown* would only be one. All of these problems should be carefully considered before a corpus is tokenised and indexed. Indexing is highly useful as a procedure to make data retrieval much faster, and may be indispensable in the case of big data. However, it should always be kept in mind that tokenisation will contain errors, which in some cases may have a significant impact on the results.

Responsible use of sources

Despite the surge in the availability of corpora over the past few decades, many pitfalls remain that do not always have an easy solution. In some cases, best practices may be formulated to avoid them. In others the best one can do is raise awareness about them. The most obvious problem remains the fact that historical data are full of holes, and we can only make 'the best use of bad data' (as per Labov's infamous quote [1994: 11]). Though the scholarly community tends to pay lip service to this problem, there are still many studies that in practice ignore or underplay it or take it into account but fail to communicate it clearly. What is even less known is that the problem is aggravated by certain unintended effects of the way corpora have been compiled or are being used side by side.

A case in point is Gries and Hilpert (2008), who argue that periodisation in historical linguistics should be established bottom-up rather than top-down. This is a very valid point, and it is worthwhile to look for an alternative bottom-up method of periodisation. As a rule, corpus compilers decide on the number and shape of periods on the basis of general historical and literary knowledge, which is more of a top-down approach. To illustrate their approach, Gries and Hilpert present a number of case studies in which they apply variability-based neighbour clustering (see Section 'Analysing data', below, for more information) to diachronic data. However, the results of their bottom-up algorithm are potentially at risk because of the nature of their sources. For their first case study, which is on the auxiliary *shall*, a significant period break was found around 1710. While this is in agreement with the customary boundary between Early Modern and Late Modern English, it is also the point where the Early Modern English corpus they used stops and the Late Modern English one (CLMET, De Smet 2005) begins. It is quite likely that rather than any inherent breaking point, it is the considerable difference in genre distribution between these corpora that is responsible for the algorithm spotting a rupture. Their second case study concerns, among other things, the decline of the prefix *ge-* in the present perfect. A steep fall is noticed around 1150, which is in line with the traditional periodisation between Old and Middle English. And yet again these results may be an artefact of the sources. While Late Old English texts in the YCOE are predominantly from the West Saxon dialect, Early Middle English texts in PPCME2 are predominantly from the

West Midlands – these problems come in addition to the data sparsity mentioned earlier. The rather abrupt loss of *ge-* therefore may mean that what is going on is a dialect shift rather than a diachronic development. Additionally, it is not clear from the paper’s methodology section whether all orthographic variants of the prefix, including the weak variants (so *i-*, *y-* next to *ge-*, *ȝe-*) were queried. Maximum clarity on what is included in a query and what not is another example of best practice that is all too commonly neglected in many studies. We want to stress that neither the method used nor the researchers’ quality is questioned here. Indeed, in a recent paper (Perek and Hilpert 2017), the method is successfully applied on a more homogenous corpus (COHA). This new paper also extends the analysis with a number of other promising data-driven methods including vector space models (for which see Section ‘Analysing data’, below.).

The impact of the sources on the analysis may be further illustrated by another loss that occurred between Old and Middle English, namely that of *weorðan* ‘be, become’, in Old English a frequent change-of-state copula and competitor of *be* in passive constructions. Figure 18.1 shows how the frequency of *weorðan* declines on the basis of two datasets. The first, represented by the full black line with circle symbols, is based on normalised frequencies (per million words) calculated on the basis of the combination of YCOE and PPCME2 (the same corpus combination as that in Gries and Hilpert 2008). The dashed line with triangle symbols shows the frequencies based on LEON 0.3 (Petré 2013b), which was specifically designed to increase comparability between Old and Middle English. There

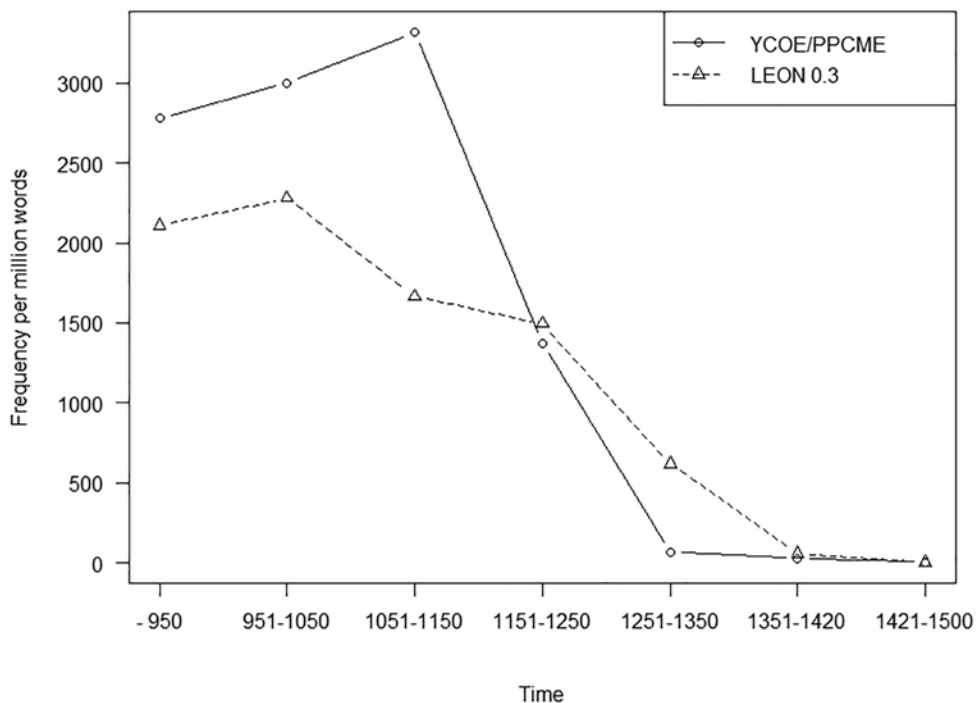


Figure 18.1 Frequency history of ‘*weorðan*’ in two different corpora

is a marked difference, with the second line going consistently down, but at a much more gradual pace than the sudden drop between OE and ME visible in the first.

A further complication relates to the normalisation of frequencies. The most common way of doing this is by taking occurrence per fixed number of words (typically 100,000 or a million). However, this method does not always yield accurate results. English has become more analytic across time, replacing, for example, case endings by prepositions or verb inflection by auxiliaries. The result of this development is that the same text will have more words than it had a millennium ago. A more accurate way of normalising therefore would be to take the number of instances per fixed number of all occurrences of the broader category to which your case study belongs. In their study on the history of the prefix *be-* (as in *become*, *befall*, etc.), for instance, Petré and Cuyckens (2008) showed that there is an increase of *be*-verb tokens between Old and Middle English. Taking lexical verb tokens as a denominator to normalise the frequencies from both periods, Middle English has 25.7 percent more *be*-verb tokens than does Old English. However, if calculated relative to the total number of tokens, this increase would only have been 16.9 percent, which is significantly less.

Another case is that of linguistic items that are related to the level of the clause. Examples are discourse markers (e.g. *truly*) or parentheticals (e.g. *I think*), which primarily have scope over the matrix clause, and are therefore not generally found in subordinate clauses. The most accurate way of calculating their frequency histories would therefore be to do this relative to the number of main clauses. Normalisation per *X* words in the corpus may, for instance, lead to underestimating any increases in the Late Middle English to Early Modern English time window, because, as has been demonstrated (cf. Green 2016: 71), the number of subordinate clauses increased significantly during this period. Such considerations, while often not really affecting the explanation for that particular case, are important if one wants to find regularities in frequency histories (such as the steepness of slopes in changing frequency rates through time). Of course, more accurate normalisations are only possible if the corpus at hand is sufficiently richly annotated.

Analysing data

In comparison to other fields of scientific inquiry such as astrophysics, palaeontology or medicine, linguistic data are relatively easy to access: a good handle on scripting languages like Perl or Python or even the use of ready-made tools like AntConc (Anthony 2011) or WordSmith (Scott 2016) allow the researcher to collect historical data, provided they are electronically available. No need for cyclotrons, excavation tools or patient consent forms. As noted in the previous section, there is a downside though. Many current-day researchers are only vaguely aware of the pitfalls with which data collection in historical linguistics is rife beyond the most obvious issues such as recall decrease due to baffling spelling variation in the past, difficulties in establishing the exact date of a historical observation and problems in understanding the meaning of historical texts.

But say you have a dataset of relevant observations which you feel confident has a decent precision and, depending on your research setup, a decent recall as well. Most of the time, historical linguists will then do some manual data preparation, involving further pruning of the dataset for false positives and enriching the data by labelling them for linguistic and extra-linguistic variables. Linguistic variables are concerned with the phonological, morphosyntactic, semantic and pragmatic aspects of each datapoint. If you are interested in the change in the 3SG inflection of verbs in Early Modern English, you may want to code whether your

observation ends in a dental fricative /θ/ or an alveolar fricative /s/ (Gries and Hilpert 2010). To the extent that this phonological feature straightforwardly maps onto the graphemes <th> and <s>, respectively, this can be done semi-automatically. For syntax, automatic data labelling is much harder. It is by no means an easy task to let a computer figure out whether a clause has the verb in second position (V2) or not, as the second position is, obviously, not the second word of the clause, but the first syntactic position after a certain constituent. For English, (historical) linguists are in the fortunate position that they have syntactically parsed corpora, but the parsing is not always transparent, especially to the novice, and dependent on particular linguistic frameworks. For other languages, syntactic annotation of one's dataset can only be achieved by going manually through each individual datapoint and coding it by hand.

For semantic and pragmatic labelling, manual labelling is the norm. If you want to code your data for animacy, a semantic feature that plays an important role in a plethora of linguistic phenomena, the common practice is to check the referent of each datapoint and discern whether it is a human, an animate being, a concrete object, or an abstract entity, for instance (other people may prefer different levels of granularity or different cut-off points on the animacy scale, of course). Progress has been made, and there are databases with psychometric measurements tapping into affective value of words (used in sentiment analysis, see Pang and Lee 2008), but for historical linguistics these databases are as yet unavailable. A solution may come from the use of distributional semantics relying on semantic vectors (see Levshina 2015, Ch. 16 for an easy-access explanation of the ins and outs), but these methods are data-hungry and may consequently not be feasible with historical corpora, although advances are currently being made in the field of historical distributional semantics (Budts *forthc.*; Dubossarsky et al. 2015, 2016, 2017; Hamilton et al. 2016; Garg et al. 2018; Perek and Hilpert 2017, among others).

The upshot is that 'digital humanities' does not equal quick automated analysis, though there is a certain tendency to go for the low-hanging fruit, and some researchers seem to steer clear as much as possible from the murky waters of complex morphosyntactic and especially semantic research.

With all these caveats in mind, let us now turn to some techniques that are used to detect trends in annotated datasets. Given the different forms in which linguistic data come – they are often non-numeric, but categorical – a wide range of techniques is applied. These techniques range from purely visual to tabular, and their use often depends on the proclivity of the individual scholar. What the two modes of representation have in common, though, is that they are usually based on quantitative measures grounded in frequencies and co-occurrence of features. Quantitative analysis of frequency measures deal with different aspects (Hilpert 2013, Ch. 1). Frequencies may either pertain to absolute frequencies of the linguistic element under study, mostly normalised per 100,000 or per million words, or with type frequencies, that is, the number of lexemes that fit a certain linguistic construction. A third way is to compare functionally more or less equivalent variants of a construction and compare whether one encroaches on the other. These kinds of frequency analysis have been popular with (historical) corpus linguists since at least the 1990s, but they often did not involve inferential statistics. This has changed lately, with a marked increase in more advanced quantitative analyses between 2005 and 2010. Despite these advancements, the field is still rapidly changing, and techniques that were considered state of the art in historical linguistics in 2005 may have become mainstream in 2015, or even low-tech in 2020. The 'quantitative turn' in linguistics (Janda 2013) has had pervasive effect.

Some techniques, like hierarchical configural frequency analysis (HCFA, see Hilpert 2013: 55–66) or correspondence analysis (Glynn 2014), are more exploratory in nature and can work without strong prior hypotheses, in a bottom-up fashion, whereas others, like regression analysis (see later) are more confirmatory in nature and are eminently suited to test prior hypotheses. Techniques also range from simple bivariate chi-square tests of significance, odds ratios or correlation measures (Gries 2013, Ch. 4, §4), to more complicated clustering techniques like principal component analysis and factor analysis (see Baayen 2008, Ch. 5), to full-on generalised linear mixed models (Baayen 2008, Ch. 7; Gries 2015) or generalised additive modelling (Winter and Wieling 2016), with many diagnostics to be checked, apart from the interpretation of the test results themselves.

The purpose of this chapter is not to give a practical introduction to the techniques that can be applied to historical linguistic data, but rather to illustrate them (see also Hilpert and Gries 2016). A nice hands-on overview of the commonly used techniques is given in Levshina (2015) and, especially for English historical linguistics, Hilpert (2013).

A good case to illustrate some fairly basic-level techniques is the study by Hinrichs and Szmrecsanyi (2007) on genitive constructions in English, where they investigate interchangeable instances of *s*-genitive and *of*-genitives (*President Kennedy's courageous actions* vs. *the courageous actions of President Kennedy*). Focusing on the American data from the 1960s and the 1990s in the BROWN and F-BROWN corpora, respectively, the situation can be visualised as a mosaic plot (Figure 18.2).

As can be appreciated from this plot, the proportion of *s*-genitives diachronically rises from the 1960s to the 1990s. To establish whether the increase is significant, we can apply a χ^2 test assessing the null hypothesis that there is no association between the two variables, namely the genitive construction and the period. This yields a *p*-value < 0.001 , so that we reject this null hypothesis. Significance is concerned with the probability that the observed effect reflects actual differences in the population, but it is not concerned with how big the effect is. To get an idea of the strength of the association, we need effect size measures. For a contingency table of frequencies of categorical variables with two levels, we can, for instance calculate Cramér's *V* (also called ϕ), ranging from 0 (no effect) to 1 (maximal effect). In the case at hand, we obtain a value of 0.17, indicating a rather modest effect. An alternative is the odds ratio, which here is 2.01, meaning that the odds of having an *s*-genitive in the 1990s is double that of the odds in the 1960s.

Hinrichs and Szmrecsanyi (2007) did not restrict themselves to American English, but have British English data as well from the 1960s and the 1990s. Here *s*-genitives make up 54.53 percent of a total of 2020 genitives in the 1960s and 54.90 percent of a total of 1945 genitives in the 1990s. If we now want to consider the difference of the regional variety (U.S. vs. UK) in the analysis, we can do different things. The most sensible analysis is a multivariate analysis where we measure the 'partial' effect of the time period and the regional variety on the expression of the genitive, for instance with multiple regression (see later). Another technique is building a three-way contingency table and subjecting it to significant tests and effect size measures. Applying a χ^2 test on the three-way contingency table returns a *p*-value < 0.001 and an effect size (Cramér's *V*) of 0.09. The drawback is that we have no idea which cells in this $2 \times 2 \times 2$ table are responsible for the overall *p*-value. The contribution to the overall *p*-value of the different cells in the table can be unearthed by looking at a so-called Cohen-friendly association plot (Gries 2013); see Figure 18.3. The plot shows a number of boxes for each of the cells, the height and the shading of which

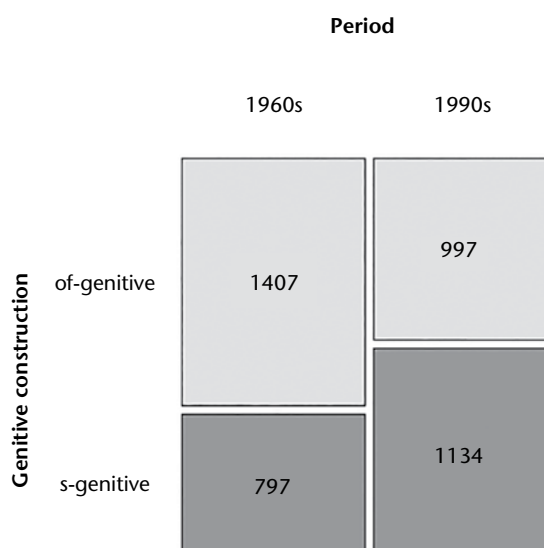


Figure 18.2 Genitive alternation in the BROWN and F-BROWN corpora

Source: Adapted from Hinrichs and Szmrecsanyi 2007

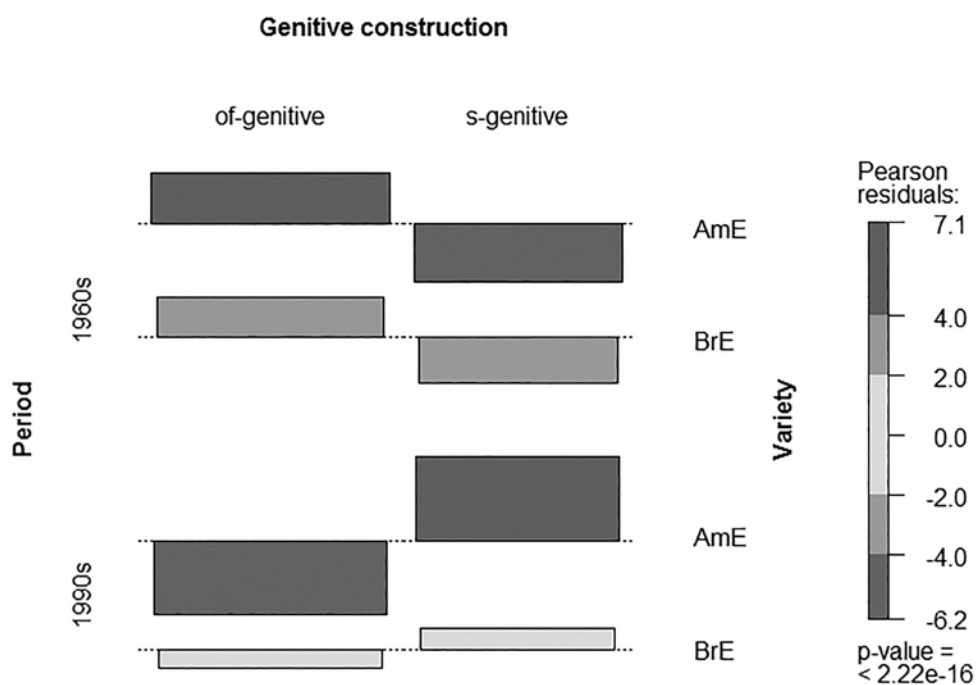


Figure 18.3 Cohen-friendly association plot of the genitive alternation in British and American English in the 1960s and the 1990s

correspond to the Pearson residuals and the width of which corresponds to the square root of the expected frequencies. Boxes above the dashed line have positive residuals, meaning that observed frequencies are higher than expected frequencies; boxes below the dashed line indicate the opposite. The intensity of the colour (redundantly) signals the absolute value of the residuals. For the genitive constructions, we see that the diachronic effect is mainly due to what happens in the United States.

Statistical tests come with their own requirements. Some assume underlying distributions, for instance normality in the population, and these are called parametric tests. Furthermore, statistical tests lose statistical power, that is, their capacity to pick up a statistical trend, when the number of observations is small, increasing the risk of type 2 errors. Small sample sizes may also affect the false-positive rate, as the test statistic may also become unreliable. For χ^2 , the rule of thumb is that the expected frequencies all must be greater than 5. And for measuring effect sizes with odds ratios in small datasets, it is sometimes advised to look at its natural logarithm, the log odds ratio (Nerbonne 2008).

By way of example, let us look at one of the findings in Petré (2015), who starts from the well-known observation that Old English has two auxiliaries for forming analytic passive preterites, namely *weorðan* and *wesan*. The semantic import of these auxiliaries is not fully overlapping, and Petré hypothesised that the seemingly free variation between *wæs ofslægen* and *weorð ofslægen*, both meaning ‘was killed’, actually reveals something about the process of killing. In line with its semantics, *weorð ofslægen* seems to have been used when the patient-subject was killed in battle while fighting and when the expression is part of the main narrative. To investigate this hypothesis, Petré looked at the cause of death underlying *ofslægen* in the early Anglo-Saxon Chronicle material (up to 950).

The contingency table underlying the mosaic plot in Figure 18.4 is amenable to a χ^2 test ($p < 0.001$), but the result is unreliable because of the low expected frequencies: the expected frequency for the ‘non-fighting’ condition for *weorðan* is 4.64 and for *wesan* 4.36, both below 5. So, we turn to Fisher’s exact test instead. What this test does is calculating the probability of the frequency distribution of the contingency table under the null hypothesis of no association between the variables, added by the sum of the probabilities of all even more extreme distributions. This test can be computationally intensive, but for a 2×2 table with a low number of observations, it is effortlessly computed on today’s laptops. The p-value here is < 0.001 . As to effect size, because one of the cells is zero, the odds ratio will either be zero or infinity, depending on how the table is set up.

In the case of the genitive constructions, recall that one of the variables was the time variable distinguishing between observations from the 1960s and observations from the 1990s, which were treated as two levels of a categorical variable. In diachronic studies, this time variable can be segmented in different ways, and is certainly not confined to a binary distinction (see later). If we have measurements in different centuries, decades or years, we can treat the variable as ordinal, or even as numeric (interval). While the latter level of measurement allows for more powerful statistics, a cautious statistician might argue that the promotion to the numeric level is injudicious, because the differences between the time periods are not necessarily equidistant. If cultural changes proceed at a non-constant speed (see e.g. Michel et al. 2011, and, specifically for lexical replacement, Bergsland and Vogt 1962), and one is interested in tracking a change against the background of cultural change, then the distance between the beginning of the sixteenth century and the end of the seventeenth century is not necessarily the same as the distance between the beginning of

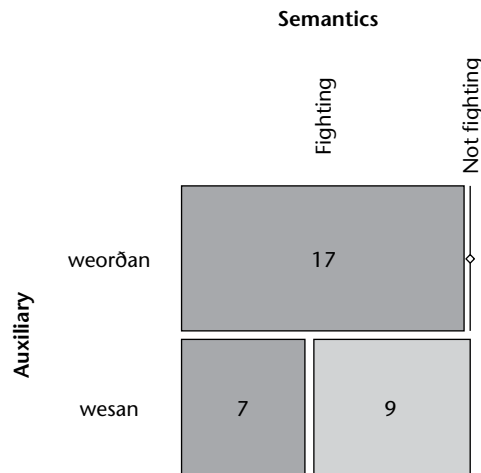


Figure 18.4 Distribution of 'weorðan' and 'wesan' as passive auxiliaries in Old English

Source: Based on Petré 2015

the nineteenth century and the end of the twentieth century. The time variable is definitely ordinal, though. The seventeenth century is inevitably in between the sixteenth century and the eighteenth century. If we want to see changes in a categorical variable over time then, we could apply statistical tests that consider the ordinal information of the time variable. A straightforward idea is to apply (non-parametric) correlation measures, such as Kendall's τ . Correlation measures measure a linear association between two numeric variables, so in principle, the linguistic outcome variable that is hypothesised to depend on the ordinal time variable should be ordinal as well. Sometimes, this is the case. Say you have three grammaticalisation stages of development of a morpheme:

- 1 A free nominal morpheme: Proto-Germanic **līka-* 'appearance, form, (dead) body', as in Gothic *allata leik þein riqizein wairþip* ('your whole body becomes dark')
- 2 A bound derivational morpheme: *-līko* 'in the manner of', as in Old High German *mannliko* ('manly')
- 3 A derivational/inflectional suffix used for adverbial function, as in Present-day English *badly*

In that case historical observations can (in principle) be labelled as one of the three stages and then treated as an ordinal variable, and the correlation can be measured on the 1-2-3 scale and the time variable (through the centuries). In many cases, however, there is no inherent order in the linguistic outcome variable. Parts of speech (verb, noun, preposition, pronoun) are hard to conceive as representing an ordinal scale, for instance. However, if the categorical variable consists of two (and only two) levels, it can be treated as ordinal. The reason is that correlation strength is unaffected by the order of the levels and only the sign will change, so permutation of the levels gives the same result. An example of this use of a

correlation measure is taken from Ghesquière and Van de Velde (2011), who set out to investigate the diachrony of *such*. One of the tendencies they lay bare is that there is a diachronic increase in the use of an accompanying adjectival modifier when *such* has intensifying meaning. The tendency is made visually clear in Figure 18.5. Gauging the non-parametric correlation between the period and the construction gives a Kendall τ of 0.50 ($p < 0.001$).

An alternative technique is a χ^2 test for the trend in proportions (Dalgaard 2008: 150). This test looks for a non-zero, linear-weighted regression slope through the proportions in each group (here: each period). The weight is defined as the number of observations for each column divided by the odds of the proportion of the ‘success’ level (whichever outcome level you want to define as the success level). For the data in Figure 18.5, this gives a p-value < 0.001 . The drawback of this technique is that we do not get an idea of the effect size, though. We only get an indication of the significant nature of the trend.

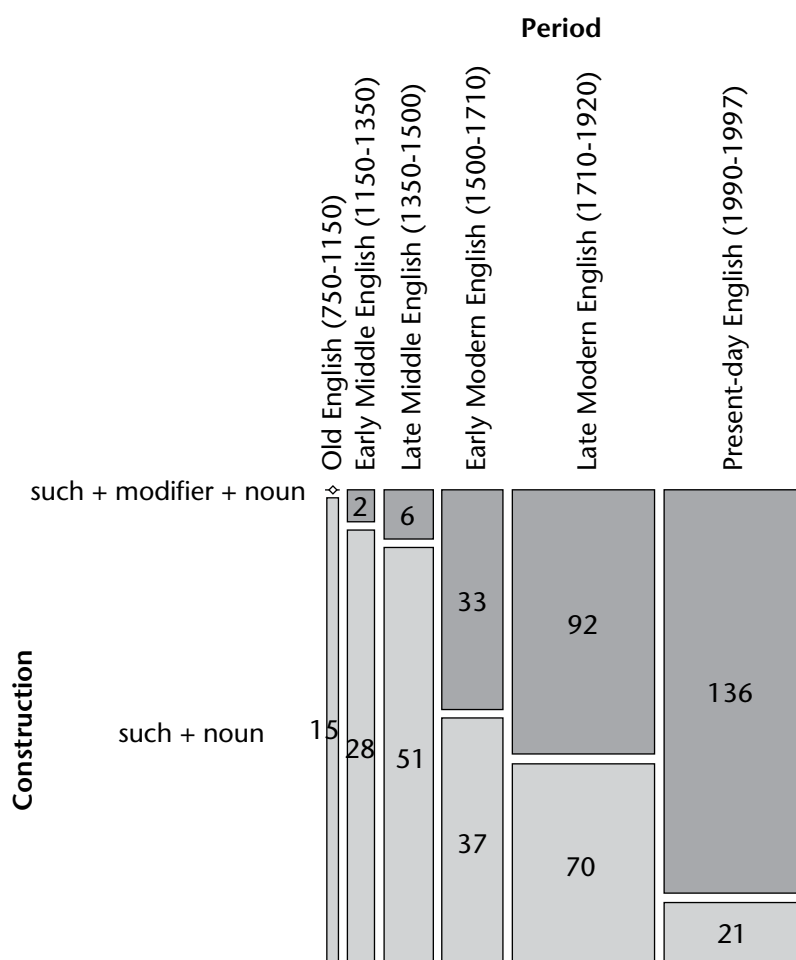


Figure 18.5 Increasing use of adjectival modifiers with intensifying ‘such’

Source: Adapted from Ghesquière and Van de Velde 2011

In an approach with more variables that interact with each other, a ‘partial correlation’ can be calculated. To illustrate this, we will have a look at a dataset from D’hoedt (2017, Ch. 6) on secondary predicate constructions of the type *I consider him an intelligent man*, with an object (*him*) and a secondary predicate XP (*an intelligent man*), without an intervening *as* preposition. The XP can take the form of an NP, AP, Adv, PP or VP. Concentrating on NPs (*I consider him an intelligent man*) and APs (*I consider him intelligent*) only, it appears that APs are gaining terrain on the NPs diachronically. This is illustrated in Figure 18.6. The Kendall τ correlation between the time variable and the syntactic status of the XP slot is 0.16 ($p < 0.001$). The construction can be used in the active voice (*I consider him intelligent*) as well as in the passive voice (*he was considered intelligent [by me]*). The use of the passive voice has diachronically receded *vis-à-vis* the active voice, though only slightly so, and there is a fair correlation between voice and XP syntax (Kendall τ 0.39, $p < 0.001$). In order to avoid the risk that the diachronic trend in the XP syntax is unduly affected by the relative rise of the active voice, or worse: an artefact of it, one might want to control for voice. This

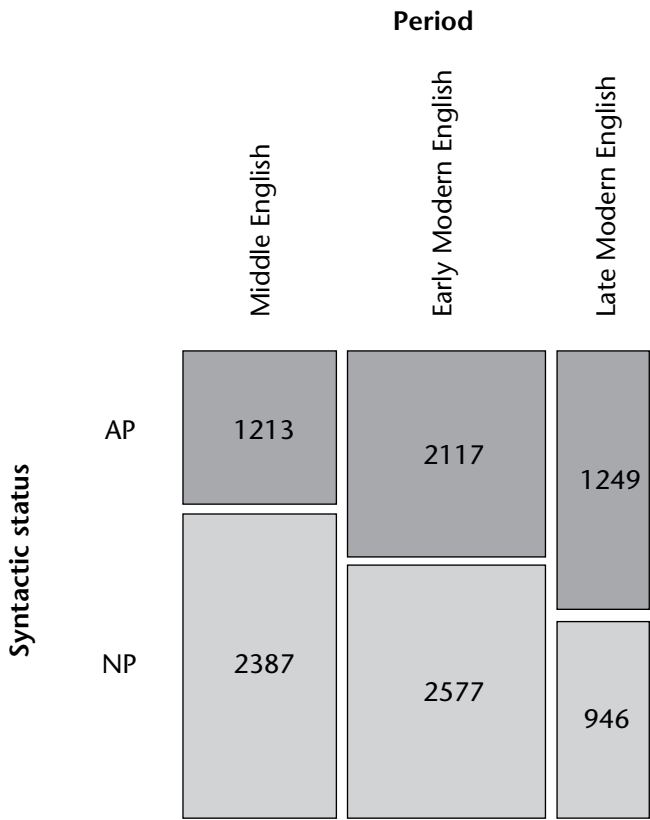


Figure 18.6 AP vs. NP in secondary predicates

Source: Adapted from D’hoedt 2017

can be done by calculating the so-called partial correlation. For the dataset at hand, Kendall τ slightly decreases, from 0.16 to 0.14 ($p < 0.001$).

Partial correlations constitute a multivariate approach, but they are fairly crude, and more sophisticated techniques are currently available. A notably popular statistical technique is regression analysis. Depending on the outcome variable, different types of regression models are needed. Many of them consist of using a link function to predict the outcome by a linear equation of a set of predictor variables, hence the name generalised linear models. For a binary outcome variable,¹ as we had in all the illustrative cases earlier, a logistic regression model is called for, using the logit (i.e. the natural logarithm of the odds), as the link function to predict the outcome Y in function of the linear equation $b_0 + \sum(b_i X_i)$ with b_0 as the intercept and b_i as the slope, X_i as the measurement of the predictor. Y is a probability (of success) here. A summation is made of several slopes ($b_1, b_2, \dots$) and regressors ($X_1, X_2, \dots$) in case the outcome is regressed on multiple predictor variables.

$$(1) \quad \text{logit}(Y) = \ln\left(\frac{Y}{1-Y}\right) = b_0 + \sum(b_i X_i)$$

The most straightforward case is when we have only one predictor, as in the case illustrated in Figure 18.6. If we treat the period as a numeric variable from 1 (Old English) to 6 (present-day English), the equation becomes:²

$$(2) \quad \text{logit}(Y) = \ln\left(\frac{Y}{1-Y}\right) = -5.18081 + 1.1422 \times \text{PERIOD}$$

So, the logit of having a modifier intervening *such* and the noun in Early Modern English for example is -0.61 . Using the inverse function of the logit, viz. the logistic function, we can turn that into a probability: 0.35. Note that this is the model's prediction. The observed proportion of the *such* + modifier + noun variant for this period is 0.47.

An extension of this generalised linear model is when the logit of the outcome variable is not defined by a linear equation, but by a polynomial. This can be illustrated by a study by Wolk et al. (2013). They look, among other things, at the diachronic development of the genitive alternation discussed earlier. Plotting the realisation of the American genitives against their periodisation yields the graph in Figure 18.7:

What we have here is a U-shaped trend, with a slight (almost imperceptible) upward tilt. This is a non-linear relationship. To model this, one can use growth curve analysis (Winter and Wieling 2016). Using a simple version (basically polynomial regression), we can add a quadratic term to the equation, so that we get:

$$(3) \quad \text{logit}(Y) = \ln\left(\frac{Y}{1-Y}\right) = b_0 + b_1 X + b_2 X^2$$

If we treat the time variable as a numeric variable (with 1650–1699 = 1, ..., 1950–1999 = 7), then the equation becomes (4). Plugging in the period number for 1851–1900 (= Period 5) yields an estimated probability of 0.19 (observed proportion is 0.20).³

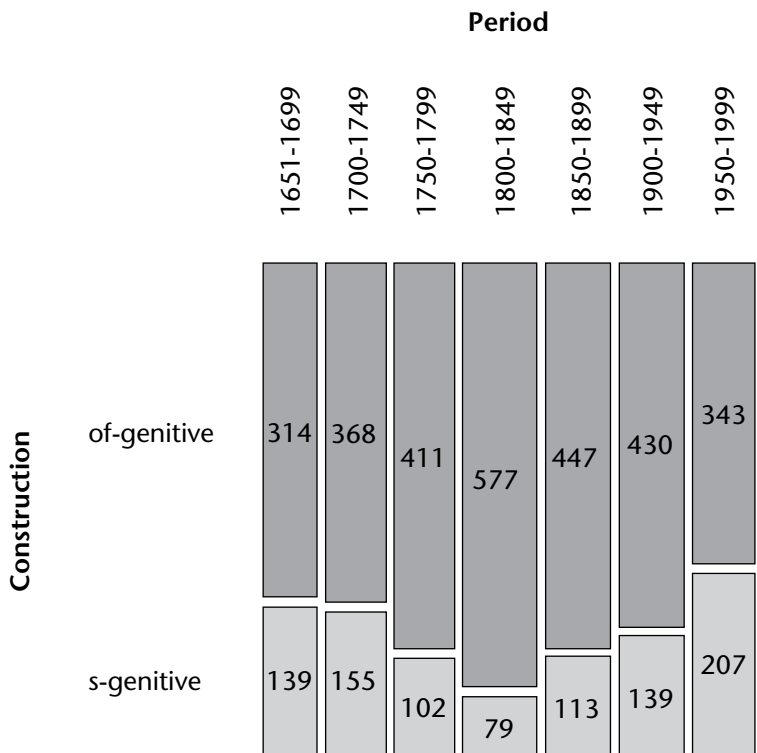


Figure 18.7 The genitive alternance
Source: Adapted from Wolk et al. 2013

(4) $\text{logit}(Y) = \ln\left(\frac{Y}{1-Y}\right) = 0.14910 - 0.88884 \times \text{PERIOD} + 0.11306 \times \text{PERIOD}^2$

Regression models prove especially useful when the outcome variable is affected by several independent variables. Such models can get complicated, especially when variables interact, and complexity is further increased by so-called mixed models allowing different intercepts and/or slopes for group levels, which can be crossed or nested. These mixed models have recently become increasingly popular (see Gries and Hilpert 2010; De Cuypere 2015; Fonteyn and Van de Pol 2016; Zimmermann 2017; Petré and Van de Velde 2018, for examples in English historical linguistics). This is not the place to discuss the technical issues involved in setting up such models, but to give an idea of what they look like, we can build a model of the aforementioned genitive alternation on the basis of Wolk et al.’s dataset. We will not replicate the full model of Wolk et al., but a trimmed-down version, where the genitive realisation is predicted by the interaction of the period and the animacy of the possessor, with random intercepts for each different possessor phrase and for each possessum phrase. The result of this mixed model is summarised in Figure 18.8. As can be seen, the probability of having an s-genitive increases over time, but only for collective, locative, and temporal

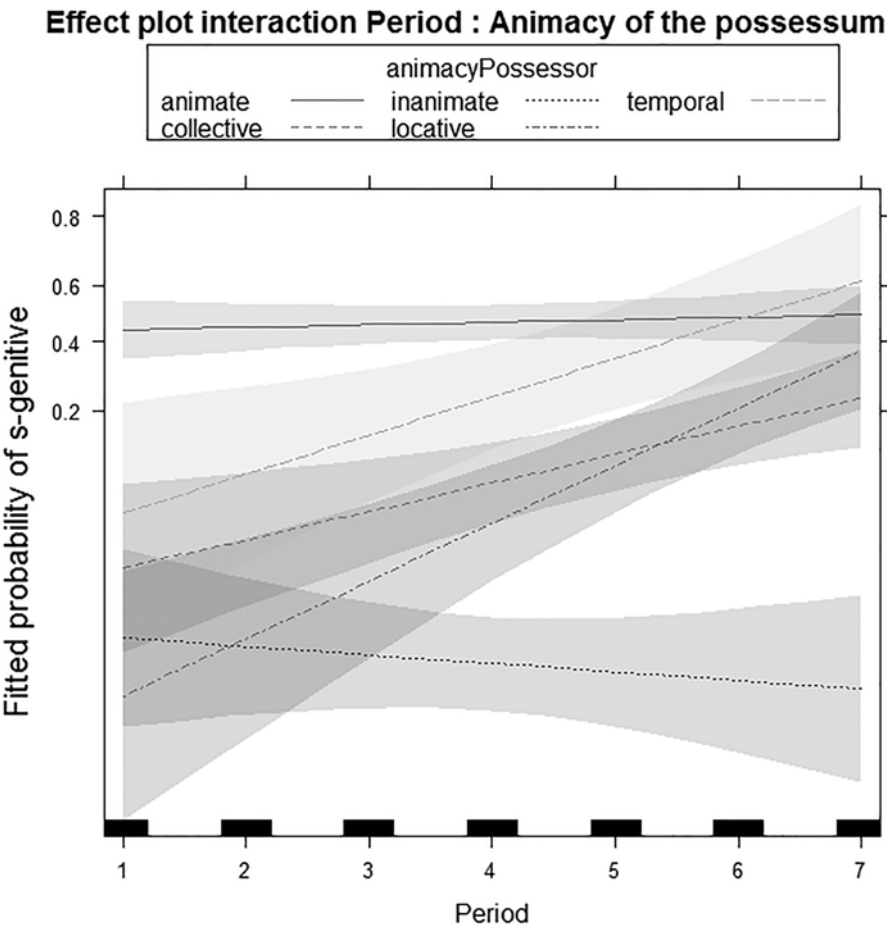


Figure 18.8 A mixed model for the diachrony genitive alternation

Source: Based on data in Wolk et al. 2013

possessors, as in (5)-(6)-(7), respectively. Only these possessor types display an increasing regression curve.

- (5) the mood of the assembly
- (6) China’s nuclear programme
- (7) yesterday’s outbreaks

Generalised linear mixed-models enjoy increasing popularity these days, but they suffer from a major drawback that not many historical linguists seem to be aware of: using a time variable as a predictor in these models, either as a main effect or in interaction with other variables, is a workable solution for observing trends, but strictly speaking it violates the assumption of independence of the datapoints. There is often a high degree of what is technically called ‘autocorrelation’ in these historical data (see Koplenig and Müller-Spitzer 2016; Koplenig 2017). This leads to an underestimation of the standard error of the time

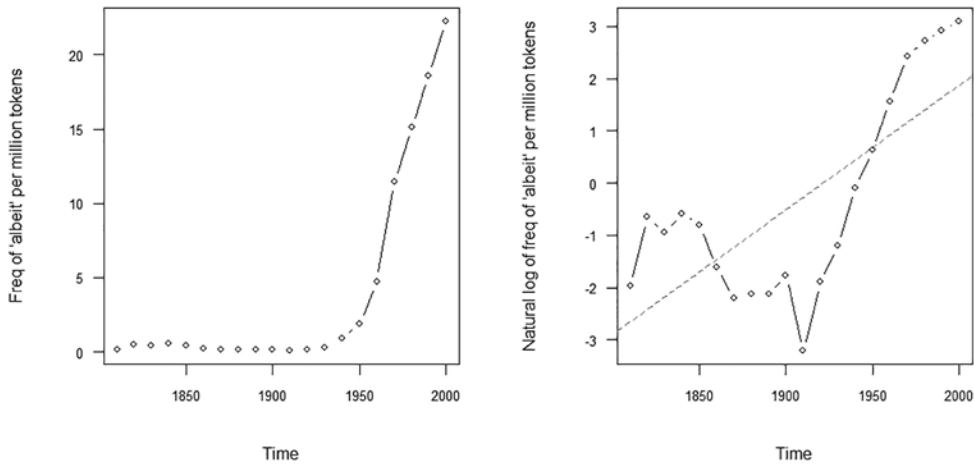


Figure 18.9 Diachrony of the conjunction *albeit* in the HANSARD corpus

predictor (Cowpertwait and Metcalfe 2009: 96). To overcome this problem, we should turn to statistical techniques that are especially tailored to deal with time, but these techniques are not common yet. One technique is time series analysis. While this technique has been applied in psycholinguistics (Baayen et al. 2018), it is new to historical linguistics (though see Koplenig and Müller-Spitzer 2016; Koplenig 2017; Rosemeyer and Van de Velde 2020). We will give a very simple case study to show how this works in practice.

The case study is about the conjunction *albeit*. Looking at the HANSARD corpus (see earlier), there appears to be a marked increase in the normalised frequency of this conjunction, that is, the frequency per million tokens per decade, see Figure 18.9, left panel. The increase looks exponential, so it makes sense to log-transform the frequency, which results in the graph in the right panel of Figure 18.9. One might be tempted to fit a linear regression through the datapoints (the dotted line), and indeed, the regression model indicates a significant effect of the time variable ($p < 0.001$), with an adjusted R^2 of 0.49 (= Pearson correlation 0.70). Though this regression model looks good, it potentially violates the independence of the datapoints: as language is transmitted through time from one speaker to the next, there is a high likelihood of dependence. To make this more concrete: the 1930 datapoint is unlikely to be totally independent from the adjacent 1920 and 1940 datapoints. This can also be seen in Figure 18.9: the residuals of adjacent time measurement points tend to cluster on the same side of the regression line. Autocorrelation can be tested with the Durbin-Watson test. In the case at hand, the test yields a p -value < 0.001 at Lag 1 (the time measurement directly adjacent).

The autocorrelation can be investigated further by visually inspecting the so-called correlogram of the observed time series and of the residual of the linear regression. The dotted lines mark the 95 percent confidence intervals. In the upper graphs in Figure 18.10, there is a significant autocorrelation at Lag 1 and Lag 2, and (negatively) at Lags 7 to 9. The lower two graphs in Figure 18.10 are the partial correlograms of the observed time series and the residuals of the regression model. Partial autocorrelation is the autocorrelation that remains after removing the effect of autocorrelation at smaller time lags. Here we see that the autocorrelation is significant at Lag 1.

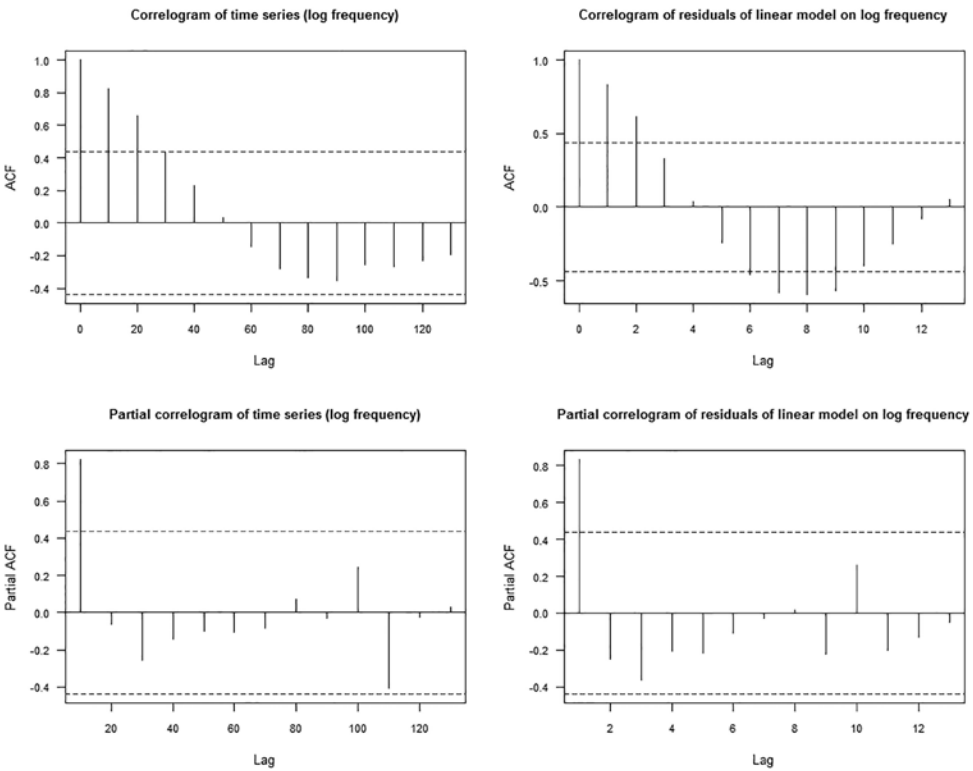


Figure 18.10 Correlograms of the diachrony of conjunction *albeit* in the HANSARD corpus

The large autocorrelation of the residuals in Lag 1 (0.8) can be taken into account in fitting a linear model using generalised least squares (Amemiya 1985; Pinheiro et al. 2017). Doing this does not really affect the regression slope here, but it increases the standard errors and the reassuringly low p-value of the original linear regression ($p < 0.001$) now increases to 0.129. In other words, not accounting for autocorrelation leads to an overly optimistic view on the linear association between time and the log-transformed frequency.

Time series analysis normally is more complex than this illustrative case. Methods have been developed for trends which show ‘seasonality’, that is, systematic variation independent of a general trend. Local temperature, for instance, may show an upward or downward trend depending on climate change, but will naturally also show systematic variation depending on month and time of the day, with nights being consistently colder than days. The latter two sources are seasonal components of the time series. Time series are analysed by decomposing them into a seasonal part, a trend part and a random part (accounting for the residuals). For the time series in Figure 18.9, there is no seasonality, so decomposition is not relevant.

Analysing time series normally consists of taking out the trend and the seasonality of the observed trend, leaving us with the random component. The correlation in the random component of the de-seasonalised and de-trended time series can then be considered for forecasting, the main objective of time series analysis. A crucial concept here is ‘stationarity’. A series is stationary if the mean, variance and the covariance of consecutive time points are not a function of time. A working time series model assumes stationarity, so any non-stationary

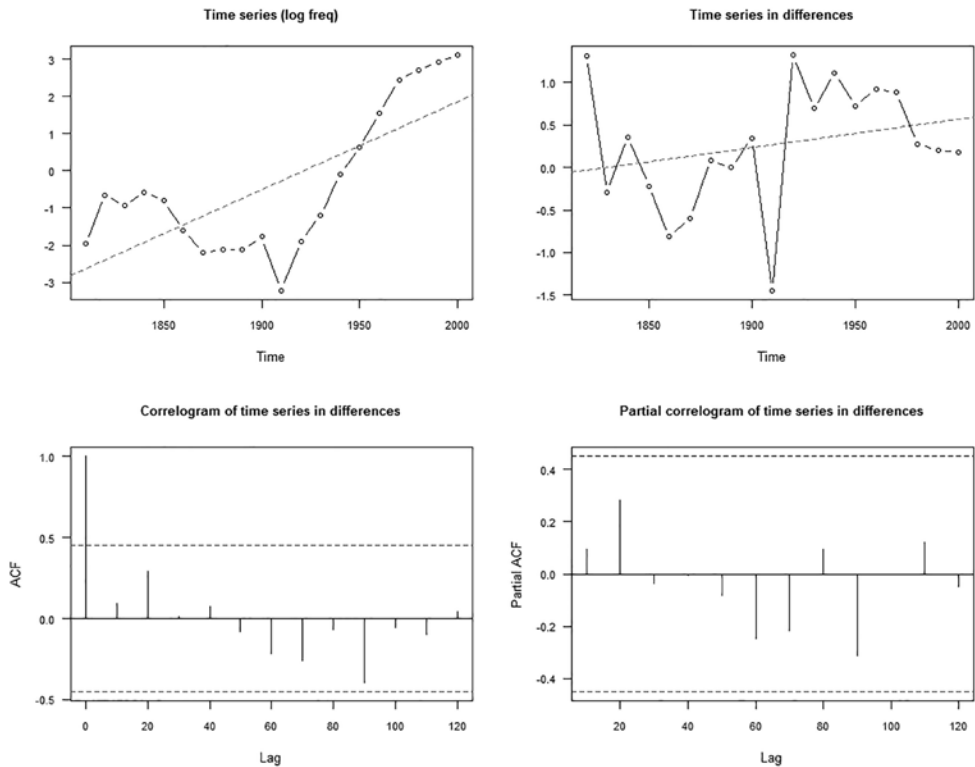


Figure 18.11 Detrending the time series of conjunction *albeit* in the HANSARD corpus by working in differences

time series first needs to be made stationary, by detrending. Detrending the time series of *albeit* in HANSARD can be achieved by working in differences. In Figure 18.11, the original time series of the log-transformed frequency (upper left panel) is contrasted with the time series in differences (upper right panel). The regression line is now flattened, and indeed, the slope is not significantly different from 0 anymore ($p = 0.284$, Adjusted $R^2 = 0.01$).

The question now is whether the time series in differences is stationary. Using Hansen's covariate-augmented Dickey Fuller test we can reject the null hypothesis of the presence of a unit root meaning we can assume stationarity ($p = 0.031$).

Finely tuned time series analysis can be done with so-called autoregressive integrated moving average (ARIMA). The details need not concern us here (see Cowpertwait and Metcalfe 2009, Ch. 7). For the simple case at hand, ARIMA modelling is definitely an overkill, but it can be done in principle. Selecting the best model for the time series of *albeit* confirms selecting the series in differences, with neither an autoregressive component nor a moving average [ARIMA(0,1,0)]. The residuals of this model are not autocorrelated (Ljung-Box test for autocorrelation $p = 0.709$). This model can be used for forecasting, as illustrated in Figure 18.12, where the dark ribbon marks the 80 percent confidence interval, and the light ribbon marks the 95 percent confidence interval.

The statistical techniques described here all have general applicability, and most of them have been developed outside of linguistics. Time series analysis, for instance, is a commonly

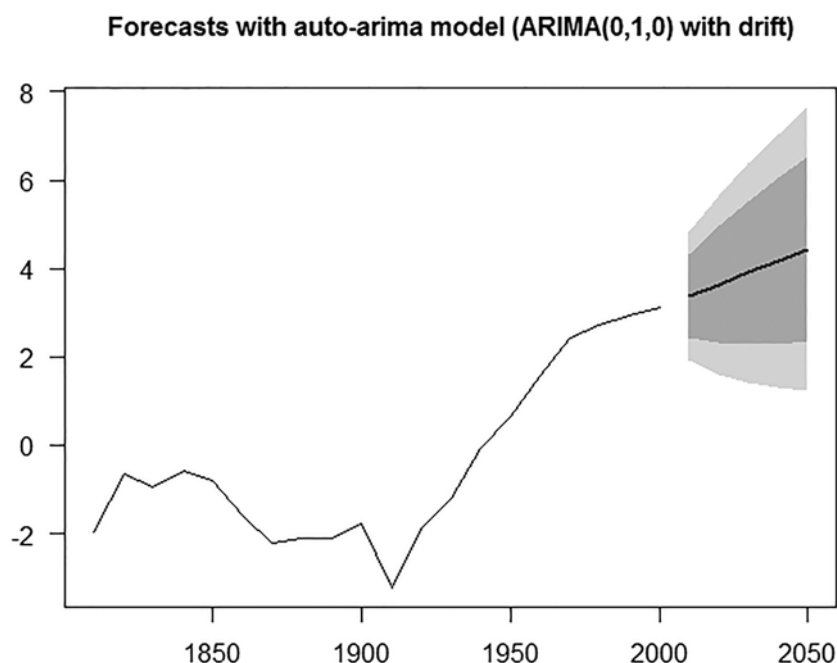


Figure 18.12 Forecasting the diachrony of *albeit* on the basis of the ARIMA model

used technique in econometrics, and mixed-model regression is popular in psychology, and – unsurprisingly – was used in psycholinguistics before it made its entry into historical linguistics. There are, however, also techniques that are especially geared to corpus linguistics and have been developed by linguists. These techniques can be seen as exemplary for digital humanities, then.

A particularly popular approach is a family of techniques that goes under the name of collocation analysis (Stefanowitsch and Gries 2003; Gries and Stefanowitsch 2004; Flach 2017). These techniques quantify the attraction (or repulsion) of lexemes and grammatical constructions, answering questions like what verbs are more likely to occur with the indirect object (*He gave/handed/offered/ . . . Mary a book*) as opposed to verbs preferring the prepositional dative (*He gave/handed/offered/ . . . a book to Mary*)? Verbs typically do not fall categorically into one class or the other, but rather have a probabilistic preference for either of the constructions. This preference can be quantified, and this is what collocation analysis aims at. Within the collocation family, there is a diachronic sibling: diachronic distinctive collexeme analysis (Hilpert 2006). This technique traces the changes in the diachronic collocational preferences in constructions and reduces these lexical preferences to broader semantic shifts. For reasons of space, we will not discuss the technique in full here, but refer the reader to Hilpert 2006, and point out that the technique is not uncontroversial (see Stefanowitsch 2006; Bybee 2010; Schmid and Küchenhoff 2013), though it has been applied with apparently good results (Hilpert 2008; Coussé 2014).

Another technique that has been designed by linguists is variability-based neighbour clustering (VNC, Gries and Hilpert 2008, 2012; Hilpert 2013: 32–45). Its goal is to arrive

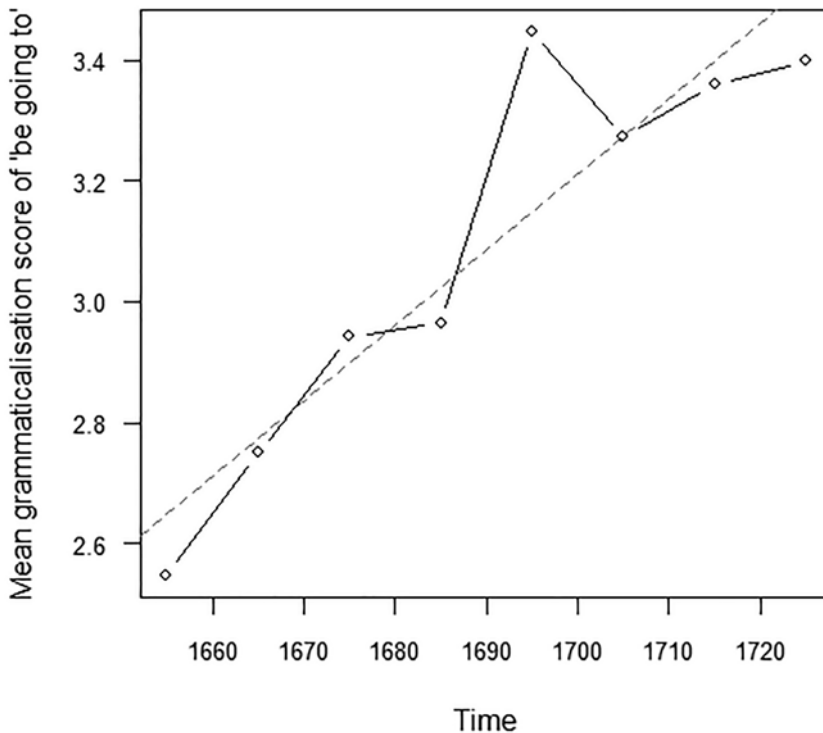


Figure 18.13 Increased grammaticalisation of 'be going to' in Early Modern English

at a bottom-up segmentation of the time variable by using a clustering method on the variability of the measurement for each measurement in time. In most diachronic studies, both inside linguistics and elsewhere, the tacit assumption is that time should be partitioned in equidistant periods, at the finest granularity level at which the outcome variable is still reasonably robustly observed. Language, however, does not change at a steady pace in time, which is part of the reason why treating the time variable as numeric is dangerous (see earlier), and particular linguistic phenomena might require an alternative split between historically meaningful periods than the set periodisation of available historical corpora.

A VNC results in a dendrogram, which can be visually inspected. To illustrate this, we can use a subset of the data of a study on the diachrony of *be going to* in Early Modern English (Petré and Van de Velde 2018). Their total dataset consists of 1369 observations of *be going to* futures in the writings of prolific authors in Early Modern English. We will focus here on the period 1650–1725, ignoring some scattered earlier and later observations. This yields a dataset of 1356 attestations. The outcome variable here is a 'score' out of 8 with diagnostics of grammaticalisation of *be going to* as a construction for expressing the future. The more grammaticalisation characteristics a given corpus observation displays, the higher its score. If it is a passive, which is considered an important diagnostic of advanced grammaticalisation (Hopper and Traugott 2003: 89), the score goes up by 1, for

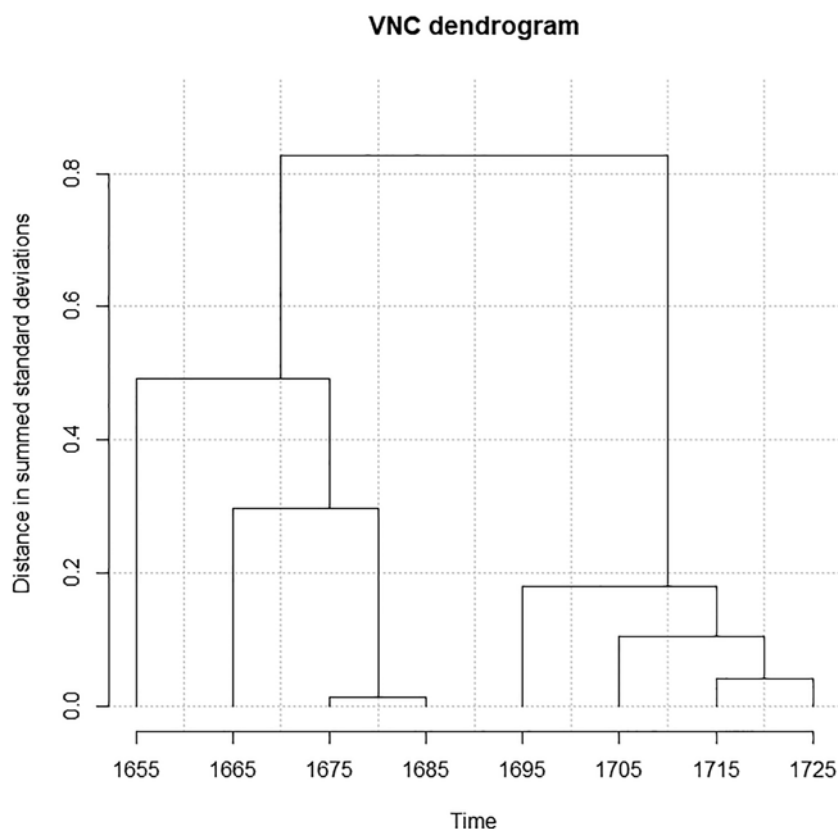


Figure 18.14 VNC analysis of increased grammaticalisation of ‘be going to’ in Early Modern English

instance. We can then aggregate the mean scores of all observations in the decades spanning from the 1650s to the 1720s. Figure 18.13 gives a view of the development through time. We see a gradual increase of the mean score, exactly as one would expect in a case of proceeding grammaticalisation. The VNC dendrogram in Figure 18.14 gives a visual impression of the optimal segmentation of the time period for this development. The highest-order split separates the pre-1690 data from the post-1690 data. In Figure 18.13, this corresponds to the leap of the curve around this time. It seems, then, that the data fall apart in two big periods that are split in the middle. The next two largest splits, above the standard deviation of 0.4 and 0.2 in the dendrogram, respectively, separate the data from the 1650s and then the data from the 1660s in the pre-1690 group. For a simple case like this one, where the time variable is measured at the fairly crude level of the decade, a VNC might not give much extra mileage in comparison to a line plot like the one in Figure 18.13, but when the curve is more erratic and the time measurement is more fine-grained, a VNC does help to reach a sensible segmentation.

Future directions

As is clear from the *tour d'horizon* in this chapter, there is no shortage of statistical techniques to analyse corpus data. And the array of techniques introduced here do not even exhaust the possibilities. Scholars often need to come up with their own techniques to answer the specific research questions in which they are interested. An example is the question to what extent aggregate tendencies at the population level can be found in individuals or generations of individual generations making up the population. This question can be tackled by separately applying common techniques introduced earlier to the textual output of different authors and comparing these authors, rather than amalgamating them in an undifferentiated corpus (see e.g. Petré 2017; Petré and Van de Velde 2018). Or the different authors can be introduced as a random effect in a generalised linear mixed model. Alternatively, scholars can come up with more specifically tailored statistical techniques.

Corpus data, however, have a severe drawback, they are messy, noisy and cannot be experimented as if in a laboratory (*pace* De Smet and Van de Velde 2017 and several papers in Hundt et al. 2017). Recently, scholars have tried to circumvent this problem by introducing methods to literally experiment on the past. A promising field is the use of computer simulations to run language change '*in silico*'. This approach is well-known in other fields of scientific inquiry, like evolutionary biology, economics and complex system science, but in linguistics, it has only recently been introduced. The majority of simulations seem to be situated in the burgeoning area of 'language evolution' (see Tallerman and Gibson 2011; Dediu and De Boer 2016 for its historiography). This fairly new field is often narrowly defined as the study of the emergence of language as such, or the transition from protolanguage to real language, and explicitly excludes the study of language change (Scott-Phillips and Kirby 2010), though others have resisted such a view and point out the similarities between the two (Steels 2017), rendering the distinction between 'phylogenetic time' and 'glossogenetic time' (Smith 2014) less severe. Simulations can take various forms. Recent years have witnessed a number of agent-based simulations to experimentally test general mechanisms of change like grammaticalisation or morphological simplification (Landsbergen et al. 2010; Beuls and Steels 2013; Lestrade 2015; Steels 2016). Such models track the evolution in speech in multi-agent generations. 'Agents' are pieces of software with pre-programmed cognitive principles and behaviour that interact with each other. While these simulations often work with invented languages, some of these studies have an architecture that stays close to existing languages, like German (Van Trijp 2013), or Dutch (Pijpops et al. 2015). Agent-based models can, in a certain sense, be seen as non-parametric models as opposed to mathematical models of language change such as Nowak et al. (2001), though some agent-based models endow their agents with mathematical formulae (Baxter et al. 2009; Blythe and Croft 2012) as well. Agent-based models have a high degree of complexity and make serious demands on the computational skills of researchers. This complexity makes it unfeasible to illustrate this approach with a concrete example in the present chapter, and we limit ourselves here to just pointing out the existence and growth of this method.

Acknowledgements

We would like to thank Benedikt Szmrecsanyi and Frauke D'hoedt for sharing their data with us.

Notes

- 1 Regression for categorical outcome variables is not restricted to binary outcome variables. See e.g. Levshina (2015, Ch.13) on multinomial models.
- 2 Treating it as a numeric variable is taking some statistical liberty. There are other solutions to deal with the categorical-ordinal nature of the period variable (e.g. Helmert coding), but these will not be discussed here.
- 3 If there is no tilt in the U-shape, the first order term of the polynomial can be dropped. We will not dig into the procedure to establish whether terms should be dropped or not. Different techniques are available for the variable selection. Strictly speaking, the variable 'Period' should first be centred (i.e. the mean should be subtracted from each observation) to avoid collinearity, but this technical issue will be glossed over here. We will also not deal with model diagnostics to see how good the fit of the model is to the data. Furthermore, we will not concern ourselves with more advanced techniques for dealing with non-linearities, like generalised additive models (GAMs), see Zuur et al. (2009, Ch. 3), Winter and Wieling (2016). This would take us too far afield in this chapter.

Further reading

- 1 Bergs, A. and Brinton, L.J. (eds.). (2012). *English historical linguistics. An international handbook* (2 vols). Berlin: De Gruyter Mouton.
Comprehensive state-of-the-art overview of historical English linguistics, including chapters on corpora and methods.
- 2 Kytö, M. and Pahta, P. (eds.). (2016). *The Cambridge handbook of English historical linguistics*. Cambridge: Cambridge University Press.
Very similar to Bergs and Brinton 2012.
- 3 Jensen, G.B. and McGillivray, B. (2017). *Quantitative historical linguistics*. Oxford: Oxford University Press.

References

- Alexander, M. and Davies, M. (2015). *Hansard corpus 1803–2005*. Available at: www.hansard-corpus.org.
- Alexander, M. and Kay, C. (2018). '... all spirits, and are melted into air, into thin air' Metaphorical connections in the history of English. In P. Petré, H. Cuyckens and F. D'Hoedt (eds.), *Sociocultural dimensions of lexis and text in the history of English*. Amsterdam: John Benjamins, pp. 61–78.
- Amemiya, T. (1985). *Advanced econometrics*. Cambridge, MA: Harvard University Press.
- Anthony, L. (2011). *AntConc* (Version 3.2.2). Tokyo, Japan: Waseda University. Available at: www.antlab.sci.waseda.ac.jp/.
- Baayen, R.H. (2008). *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge: Cambridge University Press.
- Baayen, R.H., Van Rij, J., De Cat, C. and Wood, S. (2018). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models. In D. Speelman, K. Heylen and D. Geeraerts (eds.), *Mixed-effects regression models in linguistics*. New York: Springer, pp. 49–69.
- Baxter, G.J., Blythe, R.A., Croft, W. and McKane, A.J. (2009). Modeling language change: An evaluation of Trudgill's theory of the emergence of New Zealand English. *Language Variation and Change* 21: 257–296.
- Benskin, M., Laing, M., Karaiskos, V. and Williamson, K. (2013). *An electronic version of a linguistic atlas of late mediaeval English* (eLALME). Available at: www.lel.ed.ac.uk/ihd/elalme/elalme.html.

- Bergsland, K. and Vogt, H. (1962). On the validity of glottochronology. *Current Anthropology* 3: 115–153.
- Beuls, K. and Steels, L. (2013). Agentbased models of strategies for the emergence and evolution of grammatical agreement. *PLoS One* 8(3): e58960. doi:10.1371/journal.pone.0058960.
- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing* 8(4): 243–257.
- Biber, D. and Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.
- Biber, D. and Gray, B. (2013). Being specific about historical change: The influence of sub-register. *Journal of English Linguistics* 41(2): 104–134.
- Blythe, R.A. and Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language* 88(2): 269–304.
- Broccias, C. (2008). Towards a history of English resultative constructions: The case of adjectival resultative constructions. *English Language and Linguistics* 12: 27–54.
- Budts, Sara. *Forthc.* A connectionist approach to analogy: On the modal meaning of periphrastic do in Early Modern English. *Corpus Linguistics and Linguistic Theory*.
- Bybee, J. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press.
- Coussé, E. (2014). Lexical expansion in the *have* and *be* perfect in Dutch: A constructionist prototype account. *Diachronica* 31: 159–191.
- Cowpertwait, P.S.P. and Metcalfe, A.V. (2009). *Introductory time series with R*. Dordrecht: Springer.
- Dalgaard, P. (2008). *Introductory statistics with R*. Dordrecht: Springer.
- Davies, M. (2010). *The corpus of historical American English (COHA): 400 million words, 1810–2009*. Available at: <http://corpus.byu.edu/coha/>.
- Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word corpus of historical American English. *Corpora* 7(2): 121–157.
- De Cuypere, L. (2015). The old English *to*-dative construction. *English Language and Linguistics* 19: 1–26.
- De Smet, H. (2005). A corpus of late modern English texts. *Icame Journal* 29: 69–82.
- Dediu, D. and De Boer, B. (2016). Language evolution needs its own journal. *Journal of Language Evolution* 1: 1–6.
- De Smet, H. (2006). *The corpus of late modern English texts (Extended Version)*. Department of Linguistics. Leuven: University of Leuven (CLMETEV).
- De Smet, H. (2016). How gradual change progresses: The interaction between convention and innovation. *Language Variation and Change* 28(1): 83–102.
- De Smet, H. and Van de Velde, F. (2017). Experimenting on the past: A case study on changing analysability in English *-ly*-adverbs. *English Language and Linguistics* 21(2): 317–340.
- D’hoedt, F. (2017). *Language change in constructional networks: The development of the English secondary predicate construction*. PhD Diss., KU Leuven.
- Dubossarsky, H., Grossman, E. and Weinshall, D. (2017). Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 conference on empirical methods in natural language processing*. Copenhagen: Association for Computational Linguistics, pp. 1147–1156.
- Dubossarsky, H., Tsvetkov, Y., Dyer, C. and Grossman, E. (2015). A bottom up approach to category mapping and meaning change. *NetWords 2015 Word Knowledge and Word Usage*: 66–70.
- Dubossarsky, H., Weinshall, D. and Grossman, E. (2016). Verbs change more than nouns: A bottom-up computational approach to semantic change. *Lingua e Linguaggio* 15(1): 5–25.
- Ecay, A. (2015). *The Penn-York computer-annotated corpus of a large amount of English based on the TCP (PYCCLE-TCP)*. Public release 1. Available at: <https://github.com/uoy-linguistics/pyccle>.
- Fischer, O. (2007). *Morphosyntactic change. Functional and formal perspectives*. Oxford: Oxford University Press.
- Flach, S. (2017). *Collostructions: An R implementation for the family of collostructional methods*. R package version 0.0.10. Available at: www.bit.ly/sflach.
- Fonteyn, L. and Van de Pol, N. (2016). Divide and conquer: The formation and functional dynamics of the modern English *ing*-clause network. *English Language and Linguistics* 20(2): 185–219.

- Garg, N., Schiebinger, L., Jurafsky, D. and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS*. Published ahead of print April 3, 2018. <https://doi.org/10.1073/pnas.1720347115>.
- Ghesquière, L. and Van de Velde, F. (2011). A corpus-based account of the development of English *such* and Dutch *zulk*: Identification, intensification and (inter)subjectification. *Cognitive Linguistics* 22(4): 765–797.
- Glynn, D. (2014). Correspondence analysis: Exploring data and identifying patterns. In D. Glynn and J.A. Robinson (eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy*. Amsterdam: John Benjamins, pp. 443–485.
- Green, C. (2016). *Patterns and development in the English clause system: A corpus-based grammatical overview*. Dordrecht: Springer.
- Gries, S.T. (2013). Elementary statistical testing with R. In M. Krug and J. Schlüter (eds.), *Research methods in language variation and change*. Cambridge: Cambridge University Press, pp. 361–381.
- Gries, S.T. (2015). The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10(1): 95–125.
- Gries, S.T. and Hilpert, M. (2008). The identification of stages in diachronic data: Variability-based neighbor clustering. *Corpora* 3(1): 59–81.
- Gries, S.T. and Hilpert, M. (2010). Modeling diachronic change in the third person singular: A multi-factorial, verb- and author-specific exploratory approach. *English Language and Linguistics* 14(3): 293–320.
- Gries, S.T. and Hilpert, M. (2012). Variability-based neighbor clustering: A bottom-up approach to periodization in historical linguistics. In T. Nevalainen and E.C. Traugott (eds.), *The Oxford handbook on the history of English*. Oxford: Oxford University Press, pp. 134–144.
- Gries, S.T. and Stefanowitsch, A. (2004). Extending collocation analysis: A corpus-based perspective on ‘alternations’. *International Journal of Corpus Linguistics* 9(1): 97–129.
- Hamilton, W.L., Leskovec, J. and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.
- H-CODE. *Health-corpus of diachronic English*. Available at: www.uta.edu/faculty/stvan/CADOH/CADOH-description.html.
- Hilpert, M. (2006). Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory* 2(2): 243–256.
- Hilpert, M. (2008). *Germanic future constructions: A usage-based approach to language change*. Amsterdam: John Benjamins.
- Hilpert, M. (2011). Dynamic visualizations of language change. *International Journal of Corpus Linguistics* 16(4): 435–461.
- Hilpert, M. (2013). *Constructional change in English: Developments in allomorphy, word formation, and syntax*. Cambridge: Cambridge University Press.
- Hilpert, M. and Gries, S.T. (2016). Quantitative approaches to diachronic corpus linguistics. In M. Kytö and P. Pahta (eds.), *The Cambridge handbook of English historical linguistics*. Cambridge: Cambridge University Press, pp. 36–53.
- Hinrichs, L., Smith, N. and Waibel, B. (2007). *The part-of-speech-tagged ‘Brown’ corpora: A manual of information, including pointers for successful use*. Freiburg: Department of English, Albert-Ludwigs-Universität Freiburg.
- Hinrichs, L. and Szmrecsanyi, B. (2007). Recent changes in the function and frequency of standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics* 11(3): 437–474.
- Hockey, S. (2004). The history of humanities computing. In S. Schreibman, R. Siemens and J. Unsworth (eds.), *A companion to digital humanities*. Oxford: Blackwell, pp. 3–19.
- Hoffmann, S. (2004). Using the OED quotations database as a corpus – a linguistic appraisal. *ICAME Journal* 28(4): 17–30.
- Hopper, P.J. and Traugott, E.C. (2003). *Grammaticalization*, 2nd ed. Cambridge: Cambridge University Press.

- Hundt, M. (2014). The demise of the being to V construction. *Transactions of the Philological Society* 112(2): 167–187.
- Hundt, M., Mollin, S. and Pfenninger, S.E. (eds.). (2017). *The changing English language*. Cambridge: Cambridge University Press.
- Janda, L. (ed.). (2013). *Cognitive linguistics: The quantitative turn. The essential reader*. Berlin: Mouton de Gruyter.
- Koplenig, A. (2017). Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions. *Digital Scholarship in the Humanities* 32(1): 159–168.
- Koplenig, A. and Müller-Spitzer, C. (2016). Population size predicts lexical diversity, but so does the mean sea level – why it is important to correctly account for the structure of temporal data. *PLoS One* 11(3): e0150771.
- Kroch, A. and Taylor, A. (2000). *Penn-Helsinki parsed corpus of middle English, second edition (PPCME2)*. Available at: www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-4/index.html.
- Kytö, M. and Rissanen, M. (1992). A language in transition: The Helsinki corpus of English texts. *ICAME Journal* 16: 7–27.
- Labov, W. (1994). *Principles of linguistic change I: Internal factors*. Oxford: Blackwell.
- Laing, M. and Lass, R. (2009). Shape-shifting, sound-change and the genesis of prodigal writing systems. *English Language and Linguistics* 13(1): 1–31.
- Landsbergen, F., Lachlan, R., Ten Cate, C. and Verhagen, A. (2010). A cultural evolutionary model of patterns in semantic change. *Linguistics* 48(2): 363–390.
- Lass, R. and Laing, M. (2016). Q is for what, when, where? The ‘q’spellings for OE hw. *Folia Linguistica* 37(1): 61–110.
- Leech, G. (2013). The development of ICAME and the Brown family of corpora. *Bergen Language and Linguistics Studies* 3(1). Available at: <https://bells.uib.no/index.php/bells/article/view/358/373> (Accessed 20 April 2018).
- Lestrade, S. (2015). A case of cultural evolution: The emergence of morphological case. In B. Köhnlein and J. Audring (eds.), *Linguistics in the Netherlands*. Amsterdam: John Benjamins, pp. 105–115.
- Levshina, N. (2015). *How to do linguistics with R. Data exploration and statistical analysis*. Amsterdam: John Benjamins.
- Lightfoot, D. (1979). *Principles of diachronic syntax*. Cambridge: Cambridge University Press.
- Loureiro-Porto, L. (2010). A review of early English impersonals: Evidence from necessity verbs. *English Studies* 91(6): 674–699.
- Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., The Google Books Team, Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A. and Lieberman Aiden, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science* 331(6014): 176–182. doi:10.1126/science.1199644.
- Moskovich, I. and Crespo García, B. (2007). Presenting the *Coruña Corpus*: A collection of samples for the historical study of English scientific writing. In J. Pérez Guerra, C. Jones, D. Gonzalez-Alvarez and J.L. Bueno-Alonso (eds.), *Of varying language and opposing creed’: New insights into Late Modern English*. Bern: Peter Lang, pp. 341–357.
- Nerbonne, J. (2008). *Introduction to (log) odds ratio*. Available at: www.let.rug.nl/nerbonne/teach/rema-stats-meth-seminar/presentations/Lobanova-2008-OddsRatio.pdf.
- Nevala, M. (2007). Inside and out: Address forms in 17th- and 18th-century letters. In T. Nevalainen and S.-K. Tanskanen (eds.), *Letter writing*. Amsterdam: John Benjamins, pp. 89–113.
- Nevalainen, T., Raumolin-Brunberg, H. and Mannila, H. (2011). The diffusion of language change in real time: Progressive and conservative individuals and the time depth of change. *Language Variation and Change* 23: 1–43.
- Nowak, M.A., Komarova, N.L. and Niyogi, P. (2001). Evolution of universal grammar. *Science* 291: 114–118.

- Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1–2): 1–135.
- Perek, F. and Hilpert, M. (2017). A distributional semantic approach to the periodization of change in the productivity of constructions. *International Journal of Corpus Linguistics* 22(4): 490–520.
- Pertejo, P.N. (1999). Be going to+ infinitive: Origin and development. Some relevant cases from the Helsinki corpus. *Studia Neophilologica* 71(2): 135–142.
- Petré, P. (2013a). *EEBOCorp, version 1.0*. Leuven: KU Leuven Linguistics Department [online]. Available at: lirias.kuleuven.be/handle/123456789/416330.
- Petré, P. (2013b). *LEON: Leuven English old to new, version 0.3*. Leuven: KU Leuven Linguistics Department [online]. Available at: lirias.kuleuven.be/handle/123456789/396725.
- Petré, P. (2014). *Constructions and environments: Copular, passive and related constructions in old and middle English*. Oxford Studies in the History of English 4. Oxford: Oxford University Press.
- Petré, P. (2015). What grammar reveals about sex and death: Interdisciplinary applications of corpus-based linguistics. *Digital Scholarship in the Humanities: A Journal of the Alliance of Digital Humanities Organizations* 3: 371–387.
- Petré, P. (2017). The extravagant progressive. An experimental corpus study on the history of emphatic [BE Ving]. *English Language and Linguistics* 21(2): 227–250.
- Petré, P. and Cuyckens, H. (2008). Bedusted, yet not beheaded: The role of be-’s constructional properties in its conservation. In A. Bergs and G. Diewald (eds.), *Constructions and language change*. Trends in Linguistics. Studies and Monographs, 194. Berlin: Mouton de Gruyter, pp. 133–170.
- Petré, P., Strik, O., Anthonissen, L., Budts, S., Manjavacas, E., Silva, E.-L., Standing, W. and Strik, O.A.O. (2019). Early modern multiloquent authors (EMMA): Designing a large-scale corpus of individuals’ languages. *ICAME Journal* 43: 83–122.
- Petré, P. and Van de Velde, F. (2018). The real-time dynamics of the individual and the community in grammaticalization. *Language* 94(4): 867–901.
- Pijpops D., Beuls, K. and Van de Velde, F. (2015). The rise of the verbal weak inflection in Germanic. An agent-based model. *Computational Linguistics in the Netherlands Journal* 5: 81–102.
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team. (2017). *NLME: Linear and nonlinear mixed effects models. R package version 3*. 1–131. Available at: <https://CRAN.R-project.org/package=nlme>.
- Raumolin-Brunberg, H. (2017). Sociolinguistics. In A. Bergs and L. Brinton (eds.), *The history of English. Volume 4: Early modern English*. Berlin: Mouton de Gruyter, pp. 188–208.
- Rosemeyer, M. and Van de Velde, F. (2020). On cause and correlation in language change. Word order and clefting in Brazilian Portuguese. *Language Dynamics and Change* 10.
- Roux, E. (ed.). (2010). *Two middle English translations of friar Laurent’s ‘Somme le roi’: Critical edition*. Turnhout: Brepols.
- Schmid, H.-J. and Küchenhoff, H. (2013). Collostructional analysis and other ways of measuring lexico-grammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3): 531–577.
- Scott, M. (2016). *WordSmith tools version 7*. Stroud: Lexical Analysis Software.
- Scott-Phillips, T.C. and Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences* 14: 411–417.
- Smith, A.D.M. (2014). Models of language evolution and change. *Wiley Interdisciplinary Reviews: Cognitive Science*. doi:10.1002/wcs.1285.
- Steels, L. (2016). Agent-based models for the emergence and evolution of grammar. *Philosophical Transactions of the Royal Society B* 371: 20150447.
- Steels, L. (2017). Do languages evolve or merely change? *Journal of Neurolinguistics* 43: 199–203.
- Stefanowitsch, A. (2006). Distinctive collexeme analysis and diachrony: A comment. *Corpus Linguistics and Linguistic Theory* 2: 257–262.
- Stefanowitsch, A. and Gries, S.T. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2): 209–243.

- Stenbrenden, G.F. (2015). The development of early middle English ó: Spelling evidence. In D.T.T. Haug (ed.), *Historical linguistics 2013: Selected papers from the 21st international conference on historical linguistics, Oslo, 5–9 August 2013*. Amsterdam: John Benjamins, pp. 41–51.
- Taavitsainen, I., Pahta, P., Hiltunen, T., Mäkinen, M., Marttila, V., Ratia, M., Suhr, C. and Tyrkkö, J. (eds.). (2010). *Early modern English medical texts*. CD-ROM. Amsterdam: John Benjamins.
- Taavitsainen, I., Pahta, P. and Mäkinen, M. (eds.). (2005). *Middle English medical texts*. CD-ROM. Amsterdam: John Benjamins.
- Tallerman, M. and Gibson, K.R. (eds.). (2011). *The Oxford handbook of language evolution*. Oxford: Oxford University Press.
- Taylor, A., Marcus, M. and Santorini, B. (2003). The Penn treebank: An overview. In A. Abeillé (ed.), *Treebanks: building and using parsed corpora*. Dordrecht: Springer, pp. 5–22.
- Taylor, A., Nurmi, A., Warner, A., Pintzuk, S. and Nevalainen, T. (2006). *The parsed corpus of early English correspondence. Compiled by the CEEC project team*. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive. (PCEEC).
- Taylor, A., Warner, A., Pintzuk, S. and Beths, F. (2003). *The York-Toronto-Helsinki parsed corpus of old English prose*. Electronic texts and manuals available from the Oxford Text Archive. (YPC).
- Trousdale, G. (2008). Words and constructions in grammaticalization: The end of the English impersonal construction. In S. Fitzmaurice and D. Minkova (eds.), *Studies in the History of the English language IV: Empirical and analytical advances in the study of English language change*. Berlin: Mouton de Gruyter, pp. 301–326.
- Van Trijp, R. (2013). Linguistic selection criteria for explaining language change: A case study on syncretism in German definite articles. *Language Dynamics and Change* 3(1): 105–132.
- Visser, F.T. (1963–1973). *An historical syntax of the English language*. Leiden: Brill.
- Winter, B. and Wieling, M. (2016). How to analyze linguistic change using mixed models, growth curve analysis and generalized additive modeling. *Journal of Language Evolution* 1: 7–18.
- Wolk, C., Bresnan, J., Rosenbach, A. and Szmrecsanyi, B. (2013). Dative and genitive variability in late modern English: Exploring cross-constructional variation and change. *Diachronica* 30(3): 382–419.
- Yáñez-Bouza, N. (2011). ARCHER past and present (1990–2010). *ICAME Journal* 35: 205–236.
- Zimmermann, R. (2016). *Parsed corpus of middle English verse* (PCMEP). Available at: www.pcmep.net/.
- Zimmermann, R. (2017). *Formal and quantitative approaches to the study of syntactic change: Three case studies from the history of English*. PhD Diss., Université de Genève.
- Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A. and Smith, G.M. (eds.). (2009). *Mixed effects models and extensions in ecology with R: statistics for biology and health*. New York: Springer.

Forensic linguistics

Nicci MacLeod and David Wright

Introduction

One area of applied linguistics in which there has been increasing trends in both (1) utilising technology to assist in the analysis of text and (2) scrutinising digital data through the lens of traditional linguistic and discursive analytical methods is that of forensic linguistics. Broadly defined, forensic linguistics is the application of linguistic theory and method to any point at which there is an interface between language and the law. The field is popularly viewed as comprising three main elements: (1) the (written) language of the law, (2) the language of (spoken) legal processes and (3) language analysis as evidence or as an investigative tool. The intersection between digital approaches to language analysis and forensic linguistics discussed in this chapter resides in element (3): the use of linguistic analysis as evidence or to assist in investigations. Forensic linguists might take instructions from police officers to provide investigative assistance or from solicitors for either side preparing a prosecution or defence case in advance of a criminal trial. Alternatively, they may undertake work for parties involved in civil legal disputes. Forensic linguists often appear in court to provide their expert testimony as evidence for the side by which they were retained, though it should be kept in mind that standards here are required to be much higher than they are within investigatory enterprises.

Forensic linguists find themselves confronted with a wide range of tasks, including settling matters of disputed meaning of slang terms (e.g. Grant 2017), commenting on the level of influence one speaker may have had over the contributions of another (e.g. Coulthard 1994), determining the linguistic proficiency of an individual (e.g. Cambier-Langeveld 2016) or providing an opinion on whether or not an interviewee understood the content of what her or his interviewer was saying (Brown-Blake and Chambers 2007). One of the more common linguistic problems that forensic linguists are faced with is identifying the author(s) responsible for writing a text or texts, a process widely referred to as ‘authorship analysis’. Over time, as the value that forensic linguists can bring to investigations has become clear, those in law enforcement have enlisted their assistance in the gathering of evidence and, in some cases, have provided linguist-led courses to trainee officers (see MacLeod and Grant 2017). For the most part, these activities are underpinned by scholarly attempts to refine the

methods informing such work, and standards and reliability in forensic work have found themselves at the centre of many a scholarly debate in the field (see Butters 2012).

This chapter explores the interface between forensic linguistics and digital methods in the collection and analysis of linguistic evidence, particularly in relation to questions of authorship. We describe the different processes through which digital approaches have become embedded in the forensic investigation of language use, and by drawing on two ongoing projects, we cast light on the empirical potential afforded by digital methods, tools and texts to researchers seeking to improve the standard of evidence and investigative assistance that forensic linguists can offer.

Background

Research in the field of authorship analysis has tended to concern itself with developing methods for correctly attributing authorship of lengthy literary texts to one of a small closed set of candidate authors. Addressing the question of authorship involves the comparison of the ‘anonymous’ text with known sample writings of the candidate authors, using statistical or stylistic methods, or a combination thereof. The reality of forensic casework, however, presents several challenges not addressed by the vast majority of these scholarly investigations into authorship analysis questions. First, the volume of known writing of a suspect that is available to the linguist may be rather limited, it may be of a different genre to the anonymous text and it may have been produced by different means (a desktop word processor versus a mobile phone, for example). A second set of problems are those that arise from the lack of any knowledge of base rates for linguistic features considered to be individuating, or part of an author’s *idiolectal style* (Turell and Gavalda 2013). That is, there is often insufficient appropriate data available to provide us with the general frequency of the linguistic features or patterns in any given population. Therefore, while we may be able to describe a particular linguistic choice as ‘marked’, how distinctive it actually is in a given population is arguably unknowable. We discuss proposed methods for addressing these shortcomings later in the chapter.

For the forensic authorship analyst, the detailed and systematic description and analysis of an individual’s linguistic choices when constructing texts is a point of departure. In this chapter, we are concerned with two investigative tasks into which this might feasibly feed: the attribution of anonymous texts to the correct author and the synthesis of a particular persona for investigative purposes. There has been a steady rise in the proportion of forensic linguistic cases involving digital communications (Coulthard et al. 2011: 538), ranging from cases involving disputed authorship of Short Message Service (SMS) text messages (Grant 2013) or e-mails (Turell 2010; Coulthard 2013) to ascertaining meaning in instant messaging (IM) conversations (Grant 2017). In line with this development, more scholarly attention has also been paid to refining methods for authorship analysis of electronic messages to underpin such case work (see, for example, Abbasi and Chen 2008; Koppel et al. 2002; Sousa-Silva et al. 2011). Methodological frameworks have evolved in line with requirements for them, as the work of the forensic authorship analyst has moved increasingly further into the digital realm.

Critical issues and topics

International jurisdictions impose various scientific ‘standards’ which expert evidence must satisfy in order to be admissible in court, such as the Criminal Practice Directions

Amendment No.2 [2014] in England and Wales and the Daubert Criteria in the United States. In light of these standards, the field of forensic linguistics is currently experiencing a period of intense self-reflection, and debate abounds about how best to ensure that practitioners work to the necessary scientific standards. The International Association of Forensic Linguists (IAFL) published its Code of Practice in 2013, reflecting a concern among its membership that the organisation should formally commit to upholding minimal standards in casework. It is important to keep in mind that this merely serves to provide guidelines to linguists wishing to undertake casework, and the IAFL has no enforcement capacity in relation to the Code. In this chapter we identify two challenges currently facing forensic linguistics in which digital approaches and methods can be instrumental in ensuring that standards of reliability are met and the delivery of justice improved: the rise in cyber-enabled crime and improving the reliability of analysis and evidence in forensic authorship cases. First, we outline the current climate with regard to these challenges in the field. Then, we describe two separate programmes of work being undertaken to address these challenges. Both of these case studies demonstrate how methods within digital humanities are facilitating linguistic analysis in forensic contexts that would simply not be possible without methods of data collection and analysis afforded by digital approaches. We conclude the chapter by reflecting on the digital humanist nature of the methods described and comment specifically on the computer–human relationship central to them both.

The need for digital methods: the rise in cyber-enabled crime

Sociolinguistics – of which forensic linguistics is an application – has evolved at much the same rate as other areas of the humanities in terms of the upsurge since the late 1990s in research addressing or resting upon digital data and/or methods (see, for example, Chapter 14 of this volume; Seargeant and Tagg 2014). Reflecting a trend across linguistics in general, recent scholarly interest in computer-mediated discourse has been characterised by the use of mixed methodologies, incorporating digital methods of analysis alongside those perhaps more traditionally associated with their discipline (see Bolander and Locher 2014). Individuals engage in socially meaningful interactions online in a way that typically leaves a textual trace which is easily accessible to the researcher’s scrutiny (Herring 2004). This has made such communications a site where empirical methods have been put to work shedding light on a wide range of interactive phenomena. Contemporaneously, these methods have themselves seen substantial development as software has (semi-) automated and sped up analysis in many sub-fields of the discipline. Larger, multi-million-word collections of linguistic and multimodal data can now be subject to a level of detailed inquiry not possible before, and the overall processes of collecting, storing, accessing, navigating and coding data have been accelerated immeasurably by technological advancements including corpus programmes, annotation tools, automated taggers, and qualitative analysis suites. Alongside these advances, however, a more sinister side to technology has also been evolving. Increasingly, forensic linguists are approached with digitally produced texts, and often, such as in the case of texts containing conspiracies or grooming material, the production of those digital texts is, in itself, a crime.

The UK Home Office makes a distinction between ‘pure’ cybercrime, that is, attacks on digital systems, and *cyber-enabled* crime, that is, offences traditionally committed in offline contexts which can increase in their scale or reach by being perpetrated via the Internet (McGuire and Dowling 2013). There is no doubt that child abuse is one area of criminal

activity that has been made easier and less risky by technological advances. The sexual grooming of children, that is, the preparation of children for sexual activity (Chiang and Grant 2017), is a widespread issue, and one that has escalated in line with the advancement of the World Wide Web. Increased access to large numbers of like-minded individuals *and* potential victims at the click of a button has led to figures suggesting that 60 percent of children in the UK have been sexually solicited online (Internet Watch Foundation 2013). Compounding these statistics, the anonymity afforded by the Internet means decreased levels of perceived risk involved in such activities. The rise of the dark web, a heavily encrypted and thus anonymous means of accessing online content, has rendered traditional methods of offender identification, such as tracking geolocation and Internet Protocol (IP) addresses, wholly ineffectual. Thus, online policing of child abuse has been described as being in crisis, with undercover operatives in dire need of alternative methods for pursuing the identification and prosecution of offenders.

This is a task with which the forensic linguist is well-equipped to assist. The situation highlights the need for effective, well informed, evidence-based approaches to policing the digital sphere, and methods that have been developed for authorship analysis can easily be adapted to this problem. Technological advances afford opportunities to the linguistic researcher, and thus to the investigator, just as they do the online criminal. Just as style markers are used to attribute authorship of unknown texts (see later), they are also put to good use in describing the linguistic persona of an individual (such as a victim) to the extent that that persona can be assumed by another individual (such as an undercover officer [UCO]). There is thus an obvious demand for the talents of a forensic linguist among law enforcement professionals seeking to successfully assume identities online.

Reliability of evidence

Today, most research in forensic authorship analysis is informed, either explicitly or implicitly, by the criteria for admissibility of evidence in court. This has predominantly been in the context of the United States and the Daubert Criteria (*Daubert vs. Merrell Dow Pharmaceuticals, Inc.* [1993]), which were established to ensure that expert evidence offered is ‘scientifically valid’, that is:

- 1 Whether the theory offered has been tested;
- 2 Whether it has been subjected to peer-review and publication;
- 3 The known rate of error; and
- 4 Whether the theory is generally accepted in the scientific community

(Solan and Tiersma 2004: 451)

There are a number of studies which discuss authorship methods, and the practices of forensic linguists more generally, in relation to these criteria (e.g. Solan and Tiersma 2004; Coulthard 2004; Howald 2008; Juola 2015). Similar criteria are set for expert evidence in England and Wales, the ‘reliability’ of which is assessed on the basis of factors, including the quality and completeness of the data used, the accuracy or precision rates of the method employed, the extent to which the material or method on which the expert relies has been peer-reviewed and whether the expert’s method followed established practice in the field (Crown Prosecution Service 2014). Such criteria now underpin much of the research undertaken by authorship analysts, as they reflect on how the methods they are developing would fare when tested against these parameters.

At the same time, the regulation of covert policing techniques is plagued by numerous ethical issues, making it fertile ground for scholarly endeavours by security ethicists (e.g. Nathan 2017). Central to these concerns is the potential for accusations to be levelled at UCOs of acting as *agents provocateur* (AP) or of entrapment – circumstances where a person has been induced to commit an offence which he or she would not have committed but for the inducement. The practice of assuming alternative identities in order to draw out offenders and secure an arrest is fraught with ethical difficulties in this regard, as reflected in judicial vigilance around the issue – compliance with Home Office guidelines is pivotal to the success of a prosecution. Modern approaches to establishing whether or not entrapment has taken place require consideration of whether there were reasonable grounds to suspect an offence was to be committed, as well as the police activity being properly authorised and supervised (Ellison and Morgan 2015). Thus, accurate record keeping forms an important part of preparing and carrying out an identity assumption task, and a UCO's digital pocketbook is her or his best defence against any potential allegations of impropriety. There is a place here for the forensic linguist – as we shall move on to discuss, one aspect of a victim's linguistic style that a UCO may wish to emulate is that of topic management, in a situation where any mention of sexual topics might be viewed as an enticement. If, however, a UCO is able to demonstrate that the initiation of sexual topics was a stylistic marker of the victim's linguistic persona, then such accusations may be mitigated. A thorough analysis of all facets of an individual's linguistic identity is therefore crucial in ensuring best practice and the highest chance of successful prosecutions.

Main research methods

Stylometry

Stylometry is the umbrella term used to refer to the wide range of methodological approaches to authorship analysis in which the similarity or difference between authors' styles is statistically measured on the basis of their use of a particular set of linguistic features (Coulthard et al. 2017: 153). Stylometry is used to address three different authorship problems: authorship attribution, which involves identifying which author from a set of candidates is most likely to have written an 'anonymous' text; authorship verification, which involves determining whether a particular author is responsible for a particular text; and author profiling, which is the prediction of an author's social characteristics (age, gender, personality type, occupation, ethnicity, etc.) on the basis of language use. Stylometry has developed primarily from a mathematical tradition and the most seminal early study in stylometric authorship analysis is the work of two statisticians: Fredrick Mosteller and David Wallace (1964). In an attempt to attribute a set of 12 disputed *Federalist Papers* to one of two candidate authors – Alexander Hamilton and James Madison – Mosteller and Wallace compared the disputed texts with known writings of the two authors on the basis of how frequently they used 70 high-frequency common grammatical words, including determiners, prepositions and conjunctions. They found significant differences in usage between the authors, and they assigned all 12 disputed texts to Madison, a conclusion which supports the prevailing opinion of historians. *The Federalist Papers* study was somewhat of a watershed moment for stylometric authorship analysis, and since then stylometry has benefitted greatly from the increased methodological opportunities afforded by digital and technological advances.

Increased digital capabilities have allowed for the collection of far larger datasets than before. Corpora of newspaper articles, blog posts, emails, IMs, literary texts, SMS messages,

social media posts and scientific texts, often running into the millions of words, are now regularly the objects of stylometric techniques. As for the methods themselves, they, too, have grown alongside increases in computer processing powers. Koppel et al. (2009), Stamatatos (2009) and El Bouanani and Kassou (2014) provide comprehensive surveys of the linguistic features and statistical procedures that are commonly used in stylometric authorship work. The linguistic features – or ‘style markers’ – on which similarities between texts and authors are based, range from vocabulary richness measures and function/content word frequencies to character n-grams and word sequences. Beyond lexical features, syntactic variation between authors has been captured by part-of-speech n-gram frequencies, and some studies have compared authors on the basis of semantic and pragmatic features such as synonym preferences and types of speech acts used. When it comes to the statistical means by which similarity between texts and authors is measured, Koppel et al. (2013: 318–319) distinguish between those in the ‘similarity-based’ paradigm, and those in the ‘machine learning’ paradigm, for addressing the simplest kind of authorship problem in which the true author of a document is assumed to be one of a small number of candidates. Similarity-based approaches involve the use of a given metric to measure how similar an anonymous document is to the known texts in terms of the linguistic features under investigation, and the anonymous document is attributed to that author whose known writing is most similar. In the machine learning paradigm, the set of known writings of each candidate author is used to train a classification algorithm that can assign the documents to their correct authors. This classifier is then used to estimate the author of the anonymous document.

Stylometric approaches as outlined here are a cornerstone of digital humanities (e.g. Jockers and Underwood 2016), but they have primarily been developed in non-forensic contexts. However, recent years have seen an increased focus on the forensic applications of such techniques to combat cybercrime and how they can assist in the exploration of areas, including extremist web forums (Abbasi and Chen 2005) and drug trafficking forums (Rico-Sulayes 2011). It is often held that stylometric approaches, which are tested in experimental conditions, have known error rates and involve less human interference, provide more reliable and objective results and so are best equipped to satisfy the criteria for the admissibility of evidence. However, there are some counter-arguments that stylometric approaches and their results are too difficult to present and explain to lay jurors and judges, either because there is no theoretical motivation for choosing particular style markers for analysis (Argamon and Koppel 2013: 301), because they generally require large datasets to be effective (Koppel et al. 2011) or because they are based on complex statistical assumptions with which it is unrealistic to expect the jury to engage, leading ultimately to the evidence being distorted rather than illuminated (Cheng 2013: 547). Therefore, although the digital tools are effective and well-tested, we must not lose sight of the role of the human analyst.

Stylistic approaches

Stylistic approaches contrast with stylometric approaches in that, rather than relying on computers or algorithms, it is the analysts themselves who compare the documents. Again, in the most straightforward authorship problem, this involves the expert performing a systematic linguistic analysis of the known writings available for the candidate author(s), and identifying a number of markers which they use with some level of *consistency* and which appear to be characteristic of their style(s). If a case involves more than one possible author, then they must have observably *distinctive* styles for the analysis to continue. Attention then shifts to the anonymous document(s) to identify whether the styles exhibited within them are consistent

with the known writings of any of the candidate authors. Such stylistic methods are better suited, or indeed are the only option available, when the amount of known and/or anonymous data are scant. There are also arguments that stylistic analyses are more firmly grounded in theories of linguistic variation than their stylometric counterparts (Nini and Grant 2013: 176), and that qualitative evidence is more easily demonstrable than statistical results to juries, who are more comfortable weighing up this kind of evidence (McMenamin 2002: 129; Cheng 2013: 547). However, stylistic approaches have been heavily criticised in recent years, primarily on the basis that they are too heavily reliant on the analyst's subjective judgements. Chaski (2005: 2) for instance, states that 'without the databases to ground the significance of stylistic features, the examiner's intuition about the significance of a stylistic feature can lead to methodological subjectivity and bias'. As a result, when such subjectivity is used as the foundations for stylistic contrasts drawn between texts and authors, 'these are sometimes loosely defined and can be harder to measure and evaluate' (Nini and Grant 2013: 176).

Stylometric and stylistic approaches are at best divergent, and at worst competing, with nobody knowing for sure which would fare best in any given case (Solan 2013: 557). A small number of recent studies have made promising moves in explicitly combining the stylistic and statistical (e.g. Grant 2013; Nini and Grant 2013), but one thing remains clear: stylometry has flourished as a digital humanity (see Juola 2015), while stylistic methods are yet to fully embrace the opportunities afforded to them by technological advancements. The area which can most obviously address this is the increased use of digitally collected corpora to serve as, in Chaski's (2005: 2) terms, databases in which to 'ground the significance of stylistic features', and this is demonstrated in a later section.

Analysing computer-mediated discourse

Other forensic linguistic work is concerned less with attributing or verifying authorship of texts and more with providing a descriptive account of an individual's linguistic identity such that it can be successfully assumed by another individual (see later). Such work is often designed from the outset to have direct impact on investigative training and practice. As such, the data are often tackled in such a way as to yield results that could easily be incorporated into current police training. In this case they also provide a point of departure for the development of software to semi-automate the process of analysing a linguistic persona in preparation for assuming that linguistic persona in order to engage criminals online with a view to securing an arrest. As mentioned earlier, reliability of forensic linguistic assistance offered to investigators is not required to be of the same standard as that submitted as evidence. Findings from the project outlined later are intended only to relieve some of the investigators' burden, rather than to provide evidence at a criminal trial.

In the work described in a later section, observations about experimental participants' language use through the medium of IM provide the grounding for interpretations of the relationship between language use and online identities. The focus on IM as the principle means through which these identities are discursively projected and maintained justifies the selection of computer-mediated discourse analysis (CMDA) (Herring 2004, 2012) as a starting point for the analyses. Rather than referring to an approach in itself, CMDA refers rather to an organisational principle – the fairly straightforward transferal of methods from linguistics and 'traditional' discourse analysis across to digital media. CMDA allows questions of broad social significance to be subject to robust fine-grained empirical investigation through a range of analytical methods, which simultaneously feeds into a rigorous evidence-based approach to a particular set of investigative problems.

Herring (2004) advocates the use of a ‘toolkit’ approach to CMDA, within which the researcher selects analytical tools relevant to the questions they are seeking to address. The four-level hierarchy of CMDA begins at the micro-linguistic level of structure and moves up through analyses at the level of meaning, such as scrutinising the data through the lens of speech act theory (Searle 1969, 1975), to interaction management such as the analysis of turn-taking patterns and topic control and finally to the macro-level of social phenomena. Most studies of computer-mediated discourse have tended to take phenomena at just one of these levels as their focus. In order to be maximally relevant to the forensic context, however, all four were incorporated to some degree into the analyses that formed part of the project described later. At the *structural* level choices relating to lexis, punctuation and graphology are attended to, while at the level of *meaning* turns are categorised according to the speech act(s) they perform. At the level of *interaction*, topics that are introduced, maintained or rejected by an individual are recorded, along with their preferences as far as turn length, turn structure, openings and closings are concerned. Identities, ostensibly a feature at the *social* level, are best seen as products of the resources available to them at all three of the prior linguistic levels. The linguistic nature of identity is a particularly salient consideration in online contexts, since many of the resources on which we ordinarily rely for the production and interpretation of identities tend to be absent (Donath 1999). Style markers at all levels can be used to characterise a linguistic persona, and this composite style can then theoretically be selected for performance by another individual.

As we discussed earlier, technological advances now allow for the speedy and thorough analysis of large amounts of linguistic data on an unprecedented scale. The data in this study were coded with the assistance of QSR NVivo, a piece of software designed to help organise and find rich insights into unstructured qualitative data. The application of these digital methods to often ‘messy’ data ensures that the data are easier to navigate and far less challenging to manage. Overviews can be generated of each individual’s preferences at the three linguistic levels described earlier across the duration of their IM conversations. Comparisons of these overviews then elucidate the most marked differences between participants, facilitating an examination of how successful individuals are at performing particular style markers that are not ordinarily part of their own linguistic repertoire. We expand on this process next.

Current contributions and research

We report here on two substantial research projects, both of which highlight the benefits of applying analytical methods originating from discourse analysis and sociolinguistics to substantial collections of digital texts in order to answer forensically-relevant questions. The studies also demonstrate some ways in which developments in digital methods have allowed for unprecedented exploration of large volumes of data. The works are presented here in order to showcase the breadth of research carried out under the umbrella of language as evidence, both as a subject of scholarly attention in general, but also as a means of reinforcing methods for addressing such questions in the ‘real world’. One project does this with a focus on authorship analysis and the potential for linguistic markers to distinguish individuals from one another to the extent that they can be used to assign texts of unknown authorship to the correct author. The other does so with a view to authorship synthesis, the process by which individuals incorporate features of a target persona’s language use into their own for the purposes of identity assumption.

Sample analysis

Using corpora as relevant population data: an Enron case study

While most forensic linguists have generally dismissed comparisons between their evidence and that of DNA (or fingerprinting) (e.g. Coulthard 2004: 432) primarily because of the variability of linguistic evidence, stylistic analysis stands to benefit from following some of the DNA profiling process. In particular, the use of population data can be borrowed by forensic linguists. Large-scale corpora can be used as population data to test the frequency or rarity of a particular linguistic feature or variant identified in an analysis. Such a process has the potential to address the major criticisms levelled at stylistic approaches to authorship analysis: rather than the diagnostic power of a particular style marker being determined by the analyst's intuition, population data could be used to measure the evidential value of finding a given style marker in two separate writings.

This is not a new idea. Coulthard (1994) used a corpus approach to identify the marked use of the word *then* in the famous case of Derek Bentley's disputed statement to the police. However, the corpora that he used were relatively small by modern standards: 2270 words of other witness statements and 1.5 million words of the Corpus of Spoken English (a subset of the COBUILD Bank of English). Since then, many others have championed the use of corpora in stylistic authorship analysis. One may expect the mantra of 'bigger is better' to hold in discussions of 'population' data – the more data that are available for comparison, the more reliable the tests of frequency or rarity. Therefore, the availability of general reference corpora such as the British National Corpus (100 million words), the Corpus of Contemporary American English (611 million words and growing with additions each year) and the Global Web-Based English corpus (1.9 billion words) may seem tantalising in this regard. However, the appetite amongst forensic linguists is not just for population data, but *relevant* population data. What 'relevant' means is up for debate. For some, a relevant population would comprise language users from the same 'linguistic community' (Turell and Gavaldà 2013: 499). For others, genre must be consistent with the anonymous texts under examination (Grant 2013: 473), while for Kredens (2002: 435), reliable relevant population data should be characterised by biological, social and interactional variables identical with those of the anonymous document(s) of the case. Therefore, large-scale reference corpora are not relevant enough. It is not sufficient, for instance, to compare the linguistic features exhibited in an anonymous tweet with millions of words of newspaper articles or fiction. Rather, *specialised* corpora, compiled according to consistency of entries with regard to parameters of setting, purpose, genre, discourse type and variety (Flowerdew 2004: 21) are required 'to provide population-specific statistics on usage' (Kredens and Coulthard 2012: 507).

To date, this appetite for relevant population data amongst forensic linguists has not been matched by an uptake of it. This is likely due to the time and financial cost of compiling and organising such corpora. This is especially true if the corpora are to be built for research purposes only, rather than benefitting a particular case, as the relative pay-off for the work may seem less. That is, linguists may be less inclined to expend resources on building a corpus if it is not to immediately serve their analysis in a specific case on which they are working. However, Turell (2010: 212) emphasises the role that research studies have in developing methods and assisting the expert witness, and so the construction and testing of relevant reference corpora for research purposes is an important endeavour. Coulthard (2013: 466) states that 'forensic linguists are never going to have reliable population statistics to enable

them to talk about the frequency or rarity of particular linguistic features'. Fortunately, with the development of digital humanities and related technologies, the collection, storage and analysis of specialised corpora as relevant population data is far easier than it has been in the past, and there is cause for optimism. Examples include the development of application programming interfaces (APIs) and web-crawling techniques which have made it possible to collect millions or billions of online text data such as web forum posts (e.g. Törnberg and Törnberg 2016), blogs (Schler et al. 2006) and tweets (Grieve et al. 2017). Such technologies facilitate the collection of specialised corpora of various kinds on a scale previously unimaginable and stand to greatly benefit authorship cases involving certain text types, should they be embraced by forensic linguists.

Two studies which use such corpora to demonstrate how such data can be used in an authorship context are Wright (2013) and Johnson and Wright (2014). Both studies use the Enron Email Corpus to test the frequency or rarity of authors' stylistic choices. In 2003, as part of the Federal Energy Regulatory Commission's (FERC) legal investigation into Enron's accounting practices, the email data of some Enron employees were made publicly available online. Cohen (2009) formatted this data and made it suitable for research, and it is his version of the corpus that Wright (2013) and Johnson and Wright (2014) use. In order to optimise the corpus for authorship analysis purposes, the data had to be cleaned in various ways, including the removal of duplicate emails, removal of email conversation threads, retaining only *sent* emails and the removal of unwanted metadata carried over from the original FERC set. This data clean-up was performed programmatically by David Woolls of CFL Software, Ltd. The Enron Email Corpus, which after cleaning contains 63,369 emails, sent by 176 authors totalling 2,462,151 tokens, can be considered as 'relevant population data' for a number of reasons. It is 'relevant' in that it pertains to the *linguistic community* of Enron employees, or perhaps it is more appropriate to say that it represents the communication of a specific community of practice (Eckert and McConnell-Ginet 1998: 490). Furthermore, both men and women authors are included, all of the texts in the corpus are of the same text type or genre (email), they are all written within a four year period (1998–2002) and there is some control of register with portions of the employees sharing the same occupation-related lexis. This means that when the rarity or frequency of particular style markers is being tested, Enron emails can be compared with Enron emails, and all of those in the population are writing in the same medium, in the same community of practice, working for the same company at the same time and many have the same job.

Wright (2013) focused on authors' distinctive uses of email openings and closings, which are thought to be as habitual as choices in vocabulary and syntax (Corney et al. 2001). In the first instance, the opening and closing preferences of a small subset of four male traders were identified. Some forms were common across some or all authors, such as the omission of a greeting or farewell altogether, or the use of first name only to sign off. Others, however, were found to distinguish between the four authors. For instance, one of the authors, John Arnold, used the greeting *hello*: which none of the other three did:

Extract 1: John Arnold email

<Date: Wed, 8 Nov 2000 09:07:00–0800 (PST)>
 <From: john.arnold@enron.com>
 <To: kendrick.brown@eia.doe.gov>
 <Subject:>

Hello:

I am not able to pull up the link for the short term outlook for natural gas. Can you please make sure the link is updates.

Thanks,

John

This greeting form, therefore, could be used to distinguish between this small set of authors. Transposed onto a real authorship case, this would be a style marker which characterises Arnold's style in contrast to that of his three colleagues. However, Arnold did not use this form particularly frequently; it only appeared in 7 of his 632 emails (1.11 percent). Therefore, the evidential strength of this feature as being indicative of his style is not particularly convincing. Yet, when the frequency of this feature was tested in the relevant population data (i.e. the rest of the Enron corpus), it was found to be used only once by one other author in 40,236 emails (0.002 percent). In other words, if an email in the Enron corpus contained the greeting *hello*:, it is 555 times more likely that this email belongs to Arnold than anyone else in the population (1.11 percent/0.002 percent). In Grant's (2013) terms, it is distinctive of Arnold's writing at 'population level'. Although Arnold is not a frequent user of this greeting, comparing his usage against that of a relevant population provides strong evidence of its rarity, and therefore the likelihood that an email containing that form belongs to him.

Using the Enron corpus as relevant population data in the same way, Johnson and Wright (2014) investigate the distinctiveness of collocation patterns. Launching from the assumption that collocations are unique to individuals (e.g. Hoey 2005: 181), Johnson and Wright examine the preferences of one individual, James Derrick, in the use of *please* mitigated directives. Occurring with a frequency of over 11,000, *please* is the second most key word in the Enron corpus when compared against the Corpus of Contemporary American English (*thanks* is most key, appearing 14,865 times). This indicates that the Enron dataset is representative of a community of practice in which interlocutors are generally linguistically polite to one another, and also suggests that requests or mitigated directives are this community's principal speech act in email (Johnson and Wright 2014: 47). Nevertheless, an analysis of Derrick's *please* collocates finds that his use of *please* is distinctive. First, some of his most frequently occurring *please* initial collocations are not used by any of the other Enron employees in the population. For example, there are 109 instances of *please* in his emails, 7 of which are followed by *format and*. By contrast, *please format and* is not used anywhere else in the Enron population; it is completely distinctive of his idiolectal style. Some of his preferred uses, such as *please print the*, are found with some frequency in the emails of other authors. This three-word sequence accounts for 35 of Derrick's 109 instances of *please* (32.11 percent). Besides Derrick's uses, there are 10,952 instances of *please* in the remaining Enron population, and 32 (0.29 percent) of these are part of the longer sequence *please print the*. This means that Derrick is 111 times more likely to use *please print the* than anyone else in the population (32.11 percent/0.29 percent). These findings show that even within very common communicative behaviours – *please* mitigated directives – we are able to identify distinctive patterns. Wright (2017) develops this idea further, using the 'relevant population data' approach to demonstrate that different authors in the Enron corpus, with the same job, when faced with the same communicative situation, encode and express the same speech act in different ways, and that the resultant linguistic output can be distinctive enough in the population to correctly identify or verify the author of a questions text. As some readers will have probably already noted, *please format and* and *please print the*, as

well the *hello*: greeting in Wright (2013) are rather unremarkable linguistic features. They are far from the ideal ‘smoking guns’ of authorship like a bizarre spelling of a word or a marked syntactical arrangement. Instead, these features might easily sail under the radar of forensic analysts stylistically examining an anonymous text (or a known writing). Having access to relevant population data allows us not only to identify them as potential markers of authorship, but to reveal that they are in fact *very distinctive* markers of authorship within that population and that their appearance in any two emails may be a strong indication that these emails are written by the same person.

Assuming identities online

The second study to be described in depth here set out to bring together the numerous facets of individual identity and explore the notion of idiolect in computer mediated discourse (see MacLeod and Grant 2017; Grant and MacLeod 2018, 2020). It did so from the theoretical position that identities are actively constructed for particular occasions; that they comprise a fluctuating body of resources from which individuals draw in order to conduct their social business; and that, in the online context at least, they are entirely linguistically and discursively constructed.

The research was designed with the needs of UCOs in mind, with a particular focus on investigations into online child sexual abuse as set out in the text. In this context, operatives are frequently required to assume an alternative identity online and to sustain this identity in interactions with the perpetrator in order to set up a meeting so an arrest can be made. The impetus for the work was the researchers’ prolonged involvement in the training of UCOs for this task and a realisation that both the training content and the identity assumption process itself had much to gain from rigorous academic attention and intervention. Thus, a study was carried out with the following central aims: (1) to assess what linguistic analysis is necessary and sufficient to describe an online persona so that it can be assumed by another individual and (2) develop theoretically how different individuals establish and perform their online personas through a combination of interactional, topic and stylistic choices.

The focus was on IM as a medium, a type of computer-mediated communication ‘involving two parties and done in real time (synchronously)’ (Baron 2013). Communication is facilitated through written exchanges, and, like many other types of computer-mediated communication, IM combines qualities typically associated with writing – such as lack of a visual context and paralinguistic cues, physical absence of interlocutors – with properties of spoken language, such as immediacy, informality, reduced planning and editing and rapid feedback (Georgakopoulou 2011). IM has thus been described as a ‘hybrid’ register (Tagliamonte and Denis 2008). The study draws on the logs of IM conversations between (1) online sexual groomers of children and their victims, (2) UCOs posing as individuals involved in sharing indecent images of children and their criminal targets, (3) police trainers and trainees in simulated online grooming investigations and (4) participants in experiments designed with the express purpose of investigating online linguistic identities (see Grant and MacLeod 2016; MacLeod and Grant 2016 for more on the benefits of using experimental work in research of this nature). Since they are neutral and non-triggering in content, centring on mundane topics rather than graphic descriptions of abuse, and because metadata and independent variables are so much easier to access, it is these experimentally generated IM conversations that form the basis of the discussion here. The tasks were designed with the aim of assessing what level of linguistic analysis is necessary and sufficient for the successful substitution of one interlocutor with another.

Computer-mediated communication, like all areas of language use, is multi-faceted, and a number of relevant levels of analysis can be identified as set out by Herring (2004) and described earlier. Each participant's contributions to (1) IM conversations in which they were *not* under any instruction to perform an alternative identity and (2) IM conversations in which they *had* been tasked with assuming a particular individual's identity, with varying levels of preparation, were analysed at the linguistic levels of *structure*, *meaning* and *interaction*. Thus, a detailed picture of an individual's *usual* preferences in terms of a wide range of potential style markers from the micro-level, for example, initialisms, syllable substitution, etc. (see MacLeod and Grant 2012 for the preliminary version of the taxonomy that informed this analysis) through to patterns of turn-taking and speech acts, was recorded as well as the extent to which these choices changed when they were tasked with impersonating someone else. The IM logs were not the only set of data to emerge from the experimental work. While at any one time some participants were tasked with taking on alternative linguistic identities, others were tasked with spotting when their interlocutor in an IM conversation was replaced by someone else impersonating them. Information that was noted down by participants as they prepared for the identity assumption task were available for scrutiny, as were participants' musings on why they believed the individual they were talking to was not who they implicitly claimed to be. These documents, alongside the IM logs themselves, formed an important part of the analyses.

The overwhelming majority of participants' comments about what led them to believe their interlocutor had been replaced related to features at the structural level of language, encoded in comments such as 'reduction in dots to mark ellipsis; inclusion of subject; increase in emoticons; more non verbal stylisation' and 'you changed to 'u'', are what gave people away. The second highest total represented comments relating to the interactional level. Comments in this category relate to changes in speed and pace of typing, length and breaking of turns and topic management: 'shorter turns; decrease in multi turn contributions'. Closely following behind this level was the social level, which included comments related to identities and relationships, such as 'tone became less chatty, more blunt'. It is crucial to keep in mind that these types of comments are virtually impossible to disentangle from observations at the one or more of the first three linguistic levels – an individual described by a judge as 'chatty', for example, might receive this description because they produce longer or more turns than other participants (a feature at the interactional level), or an individual may be described as 'bossy' because they issue a high proportion of directives (a feature at the level of meaning). It is a fairly straightforward matter to feed this information into training UCOs – a primary focus on low-level linguistic features will stand one in good stead for the identity assumption task, since inconsistencies at this level are the ones most easily spotted by the individual one is attempting to deceive. UCOs' attentions should also be focused at the interactional and social levels, and, according to these results, minimal attention is required at the level of meaning (although see earlier). Notes made in preparation for the identity assumption task tell a similar story to these observations. Almost 70 percent of observations about a target language style were related to potential markers at the structural level.

One important outcome of this project has been a software tool that partially automates the process of linguistic analysis of an individual's identity, expediting the preparation process for UCOs tasked with identity assumption in live investigations. The tool provides a linguistic summary of each person involved in a pre-loaded IM conversation, providing numerical information about features such as their spelling and punctuation choices and

average turn length. When there are two or more variant spellings of an item, the tool provides a ratio of the relative frequencies. The tool also provides a number of free text areas in which UCOs can record their observations about speech acts and openings, among other things. One further feature of the software is an auto-stylisation function, within which a UCO can input a phrase that is then ‘translated’ into the language of the selected individual. While not intended to replace the process of human linguistic analysis and identity performance, it is hoped that the tool will relieve some of the cognitive burden on officers engaged in these activities. This project thus represents a tangible impact of linguistic research on policing practice via digital methods.

Future directions

There can be no doubt that the digital trend is set to continue unabated and that forensic linguists will continue to be called upon to provide their assistance with either digital texts, digital investigation methods or a combination of the two. Both of the case studies that we have discussed here sit firmly within the framework of digital humanities. Both approaches combine the efficient use of computer and digital technologies with effective human interpretation and communication of results – a central tenet of ‘humanities computing’ (Unsworth 2016). Although the computer is the apparatus through which the forensic analyses described here are performed, they both require careful input, interaction and explanation by the human analyst. For instance, deciding the currency to be attached to discovering that a particular style marker 500 times more likely to be used by one author than another is a task for the analyst. More obviously still, the assumption of identities described here is performed by humans, albeit with the help of developed software. In the case studies discussed earlier, technology both facilitates the analysis *and* improves the reliability of subsequent results and forensic evidence, but crucially the human remains integral to the methods.

A clear future direction for forensic linguistics is to harness the power of digital humanities in the collection of relevant population data. Kredens and Coulthard (2012) outline some of the corpora that are available, but the amount of data available remains scarce. Researchers need to be afforded the time and the funds to build carefully constructed corpora of relevant population data for digital text types. Indeed, some of the datasets collected and used by stylometrists would be a good place to start, but the wider availability of population data would be a ‘game-changer’ in authorship analysis, both in research and casework. In a purely research context, the availability and accessibility of such corpora would allow forensic linguists to test their methods of authorship analysis in lab conditions before taking them to the police or to court (see Solan 2013). From a solely linguistic perspective, such analysis would help in making new discoveries with regard to the composition of individual style and idiolects across text types. The benefits to casework are greater still, as such datasets could provide the all-important base rate knowledge required to bolster the reliability of stylistic approaches to authorship cases. Therefore, the creation and exploration of large-scale specialised corpora should be a priority for the field and is an activity which lives unequivocally under the ‘catch-all, big tent name’ (Terras 2016: 1637) of ‘digital humanities’. Further to this, forensic linguists must continue their efforts to engage with investigators as the issue of cyber-assisted crime continues to grow. Law enforcement bodies remain in need of novel approaches to policing the online sphere, and as the work discussed earlier demonstrates, there is an important role for forensic linguistics to play in such endeavours.

Further reading

- 1 Coulthard, M., Johnson, A. and Wright, D. (2017). *An introduction to forensic linguistics: Language in evidence* (2nd ed.). London: Routledge.
This core textbook in forensic linguistics provides coverage of all major areas, issues and methods in the field.

Notes

The experimental data collected with Assuming Identities Online are available via Open Access and can be found at the following link: <http://dx.doi.org/10.5255/UKDA-SN-852099>

Assuming Identities Online – description, development and ethical implications, was funded by the ESRC, project number ES/L003279/1.

References

- Abbasi, A. and Chen, H. (2005). Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems* 20(5): 67–75.
- Abbasi, A. and Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems* 26(2): 1–29.
- Argamon, S. and Koppel, M. (2013). A systemic functional approach to automated authorship analysis. *Journal of Law and Policy* 21(2): 299–316.
- Baron, N. (2013). Instant messaging. In S. Herring, D. Stein and T. Virtanen (eds.), *Pragmatics of computer-mediated communication*. Berlin: de Gruyter, pp. 135–162.
- Bolander, B. and Locher, M. (2014). Doing sociolinguistic research on computer-mediated data: A review of four methodological issues. *Discourse, Context & Media* 3: 14–26.
- Brown-Blake, C.N. and Chambers, P. (2007). The Jamaican Creole speaker in the UK justice system. *The International Journal of Speech, Language and the Law* 14(2): 269–294.
- Butters, R. (2012). Retiring president’s closing address: Ethics, best practices, and standard. In S. Tomblin, N. MacLeod, R. Sousa-Silva and M. Coulthard (eds.), *Proceedings of the international association of forensic linguists tenth biennial conference*. Birmingham: Centre for Forensic Linguistics, pp. 351–361.
- Cambier-Langeveld, T. (2016). Language analysis in the asylum procedure: A specification of the task in practice. *The International Journal of Speech, Language and the Law* 23(1): 25–41.
- Chaski, C.E. (2005). Who’s at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence* 4(1): 1–14.
- Cheng, E.K. (2013). Being pragmatic about forensic linguistics. *Journal of Law and Policy* 21(2): 541–550.
- Chiang, E. and Grant, T. (2017). Online grooming: Moves and strategies. *Language and Law/Linguagem e Direito* 4(1): 103–141.
- Cohen, W.W. (2009). *Enron email dataset* [online]. Available at: www.cs.cmu.edu/~enron/.
- Corney, M., Anderson, A., Mohay, G. and de Vel, O. (2001). *Identifying the authors of suspect Email* [online]. Available at: <http://eprints.qut.edu.au/8021/1/CompSecurityPaper.pdf>.
- Coulthard, M. (1994). On the use of corpora in the analysis of forensic texts. *The International Journal of Speech, Language and the Law (Forensic Linguistics)* 1(1): 27–43.
- Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics* 24(4): 431–447.
- Coulthard, M. (2013). On admissible linguistic evidence. *Journal of Law and Policy* 21(2): 441–466.
- Coulthard, M., Grant, T. and Kredens K. (2011). Forensic linguistics. In R. Wodak, B. Johnstone and P. Kerswill (eds.), *The SAGE handbook of sociolinguistics*. London: Sage, pp. 531–544.
- Coulthard, M., Johnson, A. and Wright, D. (2017). *An introduction to forensic linguistics: language in evidence*, 2nd ed. London: Routledge.

- Crown Prosecution Service. (2014). *Expert evidence*. Available at: www.cps.gov.uk/legal/assets/uploads/files/expert_evidence_first_edition_2014.pdf.
- Donath, J. (1999). Identity and deception in the virtual community. In M. Smith and P. Kollock (eds.), *Communities in cyberspace*. Abingdon: Routledge, pp. 29–59.
- Eckert, P. and McConnell-Ginet, S. (1998). Communities of practice: Where language, gender and power all live? In J. Coates (ed.), *Language and gender: A reader*. Oxford: Blackwell, pp. 484–494.
- El Bouanani, S.E.M. and Kassou, I. (2014). Authorship analysis studies: A survey. *International Journal of Computer Applications* 86(12): 22–29.
- Ellison, M. and Morgan, A. (2015). *Review of possible miscarriages of justice: Impact of undisclosed undercover police activity on the safety of convictions*. Report to the Attorney General.
- Flowerdew, L. (2004). The argument for using English specialized corpora to understand academic and professional settings. In U. Connor and T.A. Upton (eds.), *Discourse in the professions: Perspectives from corpus linguistics*. Amsterdam: John Benjamins, pp. 11–33.
- Georgakopoulou, A. (2011). “On for drinkies?” Email cues of participant alignments. *Language@Internet* 8.
- Grant, T. (2013). TXT 4N6: Method, consistency, and distinctiveness in the analysis of SMS text messages. *Journal of Law and Policy* 21(2): 467–494.
- Grant, T. (2017). Duppying yoots in a dog eat dog world, kmt: Determining the senses of slang terms for the courts. *Semiotica* 216: 479–495.
- Grant, T. and MacLeod, N. (2016). Assuming identities online: Linguistics applied to the policing of online paedophile activity. *Applied Linguistics* 37(1): 50–70.
- Grant, T. and MacLeod, N. (2018). Resources and constraints in linguistic identity performance – A theory of authorship. *Language and Law/Linguagem e Direito* 5: 1.
- Grant, T. and MacLeod, N. (2020). *Language and online identities: The undercover policing of internet sexual crime*. Cambridge: Cambridge University Press.
- Grieve, J., Nini, A. and Guo, D. (2017). Analyzing lexical emergence in modern American English online. *English Language and Linguistics* 21: 99–127.
- Herring, S. (2004). Computer-mediated discourse analysis: An approach to researching online behavior. In S.A. Barab, R. Kling and J.H. Gray (eds.), *Designing for virtual communities in the service of learning*. Cambridge: Cambridge University Press, pp. 338–376.
- Herring, S. (2012). Discourse in Web2.0: Familiar, reconfigured and emergent. In *Discourse 2.0: Language and new media*. Georgetown: Georgetown University Round Table on Languages and Linguistics 2011, pp. 1–29.
- Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.
- Howald, B.S. (2008). Authorship attribution under the rules of evidence: Empirical approaches – a lay-person’s legal system. *The International Journal of Speech, Language and the Law* 15(2): 219–247.
- Internet Watch Foundation. (2013). *ChildLine and the internet watch foundation form new partnership to help young people remove explicit images online* [Online]. Available at: www.iwf.org.uk/about-iwf/news/post/373-childline-and-the-internet-watch-foundation-form-new-partnership-to-help-young-people-remove-explicit-images-online.
- Jockers, M.L. and Underwood, T. (2016). Text-mining the humanities. In S. Schreibman, R. Siemens and J. Unsworth (eds.), *A new companion to the digital humanities*. London: Wiley Blackwell, pp. 291–306.
- Johnson, A. and Wright, D. (2014). Identifying idiolect in forensic authorship attribution: An n-gram textbite approach. *Language and Law/Linguagem e Direito* 1(1): 37–69.
- Juola, P. (2015). The Rowling case: A proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities* 30(suppl 1): i100–i113.
- Koppel, M., Argamon, S. and Shimoni, R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17(4): 401–412.
- Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology* 60(1): 9–26.
- Koppel, M., Schler, J. and Argamon, S. (2011). Authorship attribution in the wild. *Language Resources and Evaluation* 45(1): 83–94.

- Koppel, M., Schler, J. and Argamon, S. (2013). Authorship attribution: What's easy and what's hard? *Journal of Law and Policy* 21(2): 317–332.
- Kredens, K. (2002). Towards a corpus-based methodology of forensic authorship attribution: A comparative study of two idiolects. In B. Lewandowska-Tomaszczyk (ed.), *PALC'01: practical applications in language corpora*. Frankfurt am Main: Peter Lang, pp. 405–437.
- Kredens, K. and Coulthard, M. (2012). Corpus linguistics in authorship identification. In P. Tiersma and L.M. Solan (eds.), *The Oxford handbook of language and law*. Oxford: Oxford University Press, pp. 504–516.
- MacLeod, N. and Grant, T. (2012). Whose Tweet? Authorship analysis of micro-blogs and other short form messages. In *Proceedings of the international association of forensic linguists' 10th biennial conference*. Birmingham: Aston University, July. Available at: www.linguisticaforense.ufsc.br/tiki-index.php?page=IAFL+2011.
- MacLeod, N. and Grant, T. (2016). “You have ruined this entire experiment . . . shall we stop talking now?” Orientations to being recorded as an interactional resource. *Discourse, Context & Media* 14: 63–70.
- MacLeod, N. and Grant, T. (2017). “Go on cam but dnt be dirty”: Linguistic levels of identity assumption in undercover online operations against child sex abusers. *Language and Law/ Linguagem e Direito* 4(2): 157–175.
- McGuire, M. and Dowling, S. (2013). *Cyber crime: A review of the evidence*. Home Office Research Report 75.
- McMenamin, G.R. (2002). *Forensic linguistics: Advances in forensic stylistics*. Boca Raton, FL: CRC Press.
- Mosteller, F. and Wallace, D. (1964). *Inference and disputed authorship: The federalist*. Reading, MA: Addison-Wesley Publishing Company Inc.
- Nathan, C. (2017). Liability to deception and manipulation: The ethics of undercover policing. *Journal of Applied Philosophy* 33(3): 370–388.
- Nini, A. and Grant, T. (2013). Bridging the gap between stylistic and cognitive approaches to authorship analysis using systemic functional linguistics and multidimensional analysis. *The International Journal of Speech, Language and the Law* 20(2): 173–202.
- Rico-Sulayes, A. (2011). Statistical authorship attribution of Mexican drug trafficking online forum posts. *The International Journal of Speech, Language and the Law* 18(1): 53–74.
- Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. (2006). Effects of age and gender on blogging. In *Proceedings of 2006 AAAI Spring symposium on computational approaches for analyzing weblogs*. Available at: http://u.cs.biu.ac.il/~schlerj/schler_springsymp06.pdf.
- Sergeant, P. and Tagg, C. (eds.). (2014). *The language of social media: Identity and community on the internet*. London: Palgrave Macmillan.
- Searle, J.S. (1969). *Speech acts*. Cambridge: Cambridge University Press.
- Searle, J.S. (1975). Indirect speech acts. In P. Cole and J.L. Morgan (eds.), *Syntax and semantics 3: Speech acts*. New York: Academic Press, pp. 59–82.
- Sousa-Silva, R., Laboreiro, G., Sarmiento, L., Grant, T., Oliveira, E. and Maia, B. (2011). ‘twazn me!!! ;(’ Automatic authorship analysis of micro-blogging messages. In R. Muñoz, A. Montoyo and E. Métais (eds.), *Natural language processing and information systems. NLDB 2011. Lecture Notes in Computer Science*, vol 6716. Heidelberg: Springer, pp. 161–168.
- Solan, L. (2013). Intuition versus algorithm: The case for forensic authorship attribution. *Journal of Law and Policy* 21(2): 551–576.
- Solan, L. and Tiersma, P. (2004). Author identification in American courts. *Applied Linguistics* 25(4): 448–465.
- Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60(3): 538–556.
- Tagliamonte, S. and Denis, D. (2008). Linguistic ruin? LOL. Instant messaging and teen language. *American Speech* 83(1): 3–34.
- Terras, M. (2016). A decade in digital humanities. *Journal of Siberian Federal University: Humanities & Social Sciences* 9(7): 1637–1650.

- Törnberg, A. and Törnberg, P. (2016). Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum. *Discourse & Society* 27(4): 401–422.
- Turell, M.T. (2010). The use of textual, grammatical and sociolinguistic evidence in forensic text comparison'. *The International Journal of Speech, Language and the Law* 17(2): 211–250.
- Turell, M.T. and Gavalda, N. (2013). Towards an index of idiolectal similitude (or distance) in authorship analysis. *Journal of Law and Policy* 21(2): 495–514.
- Unsworth, J. (2016). What is humanities computing and what is it not? In M. Terras, J. Nyhan and E. Vanhoutte (eds.), *Defining digital humanities: A reader*. London: Routledge, pp. 35–48.
- Wright, D. (2013). Stylistic variation within genre conventions in the Enron email corpus: Developing a text-sensitive methodology for authorship research. *International Journal of Speech, Language and the Law* 20(1): 45–75.
- Wright, D. (2017). Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International Journal of Corpus Linguistics* 22(2): 212–241.

Corpus linguistics

Gavin Brookes and Tony McEnery

Introduction

Corpus linguistics refers to a group of methods that use specialist computer programmes to study language in large bodies of machine-readable text (McEnery and Wilson 2001). These bodies of data are known as *corpora* (singular: *corpus*, Latin for ‘body’) and are assembled with the aim of representing a language or particular linguistic variety on a broad scale. Indeed, because they use computers, studies in corpus linguistics are routinely based on datasets that are much larger than would be practical to analyse without computational assistance, with corpora often amounting to millions and sometimes billions of words. Corpora can either be general or specialised. General corpora are designed to represent a whole language at a particular point in time, for example, the British National Corpus (Aston and Burnard 1998), which represents written and spoken general British English during the 1980s–1990s. Specialised corpora, on the other hand, are designed to represent language used in specific contexts and are usually compiled with a particular project or study in mind, for example, the corpus of UK press articles about refugees and asylum seekers assembled and analysed by Baker and colleagues (Baker et al. 2008). Because they strive to represent language on a broader scale, general corpora are usually larger than specialised corpora. Once a corpus has been built or selected, the specialist computer programmes that are used in corpus linguistics allow human researchers to perform statistical analyses on their corpora, providing the opportunity to learn about the most frequent and statistically salient linguistic patterns in their data. These patterns can then be scrutinised from a more qualitative, theory-informed perspective (e.g. discourse analysis [Chapter 10] or critical discourse analysis [CDA] [Chapter 11]).

This chapter will introduce readers to the key principles and methods of corpus linguistics and both discuss and demonstrate its (potential) contribution to digital humanities research. The chapter begins by setting out the general features of corpus linguistics, including its main strengths and limitations as a collection of tools for linguistic analysis. The section on critical issues and topics reviews a series of key debates in the field, some of which have shaped the view of corpus linguistics to which we, the authors, subscribe. Following this, we will discuss corpus linguistics’ current contributions to research within

linguistics and the digital humanities, including comparing corpus techniques against other digital methods that are popular amongst humanities researchers. We will then introduce a series of established techniques in corpus linguistics and demonstrate their use and utility by means of a sample analysis of a large corpus of online patient feedback. The chapter concludes with a summary of its main arguments and discussion points and by looking to the future of corpus linguistics in the digital humanities.

Background

Corpus linguistics focuses on something of a diverse group of procedures or methods for studying language. This heterogeneity is an important observation to make since, despite several established points of departure, the field of corpus linguistics comprises a wide range of approaches. As McEnery and Hardie (2012: 1) put it, '[c]orpus linguistics is not a monolithic, consensually agreed set of methods and procedures for the exploration of language. . . . Differences exist within corpus linguistics which separate out and subcategorise varying approaches to the use of corpus data'. These discriminating factors will be discussed in the next section. However, before this, we will first consider some of the more general, widely agreed features of corpus linguistics, as these provide a good starting point for our discussion and for this chapter.

Because of their size, if assembled in a careful way, corpora can offer greater representativeness of a language or variety of interest than, say, a single or handful of texts. Representativeness, according to Biber (1993: 243), refers to the 'extent to which a sample includes the full range of variability in a population'. Biber argues that when designing a corpus, variability can be considered both from situational and linguistic perspectives. For a corpus to be representative of a target language or variety, it therefore needs to contain texts or samples of texts which represent (1) the range of text types in that language or variety and (2) the range of linguistic distributions in that language or variety (Biber 1993). The benefit of analysing representative corpora is that they usually allow analysts to claim that their findings are more generalisable with respect to the type of language they are investigating. A related strength of corpus methods is that they are useful for pinpointing what Baker (2006: 13) describes as the 'incremental effect of discourse'; that is, the propensity for discourses to be subtly established through linguistic patterns that might feature sparingly in one or two texts, but become significant when considered as part of a broader discourse type or collection of texts.

The next generalisation follows from the first; because the data are machine-readable, corpus linguists can use specialist software to search and analyse large corpora rapidly and reliably, producing analyses that are less subjective and guided by more objective criteria, such as frequency and statistical significance (McEnery and Hardie 2012). This can help analysts (particularly those adopting a critical position) to overcome the accusation that certain texts or examples from their data have been cherry-picked because they support an *a priori* argument (Widdowson 2004). A further advantage of corpus methods is that the specialist computer programmes that underpin them can perform frequency counts and complex statistical calculations faster and more reliably than the human mind, and so can reveal linguistic patterns that might otherwise evade manual observation or run counter to human intuition (Baker 2006: 10–14).

Like any other methodology, however, corpus linguistics is not free of limitations. One limitation concerns what corpora have the power to represent. The rendering of any text or collection of texts into a corpus is a transformative process, the product of which

bears important differences to the originals. Because corpora tend to contain transcriptions of spoken language or written material, modes like gesture (for example, in spoken interactions) and visuals (for example, in newspaper texts) are more difficult, though not impossible, to analyse using corpus techniques. A related criticism of corpora is that the process of converting texts into a corpus divorces them from the social contexts in which they were originally produced and understood. The corpus analyst must therefore carry out more work to ensure that their analysis is cognizant of the role that contexts of situation and culture play in the production and consumption of the texts contained in their corpus. One way this can be done is by drawing on extra-corpus theories and resources which can provide clues about the historical and sociopolitical contexts in which the texts contained in the corpus were originally situated. For example, in his study of swearing in English, McEnery (2005) coupled corpus data with theories and supporting data from beyond linguistics to gain a deeper insight into historical attitudes towards swearing. This allowed him to move from describing to *explaining* patterns of bad language use over time.

Another limitation of corpus approaches is that they tend to focus on the linguistic features that are *present* in the data, but tell us less about what is *absent*. Finally, although computer programmes can be useful for identifying frequent and statistically interesting linguistic patterns, they cannot interpret those patterns or tell us why they are significant; it is up to the human analyst to dig deeper (and often wider – drawing on corpus-external sources) to make sense of and explain the significance of the patterns in their data. Whatever the case, these criticisms should not discourage the use of corpus methods but should be considered by researchers when deciding on whether and how to use corpus approaches, as well as when working out what steps can be taken to overcome any methodological shortcomings.

Critical issues and topics

Having outlined important generalisations that can be applied across the spectrum of corpus linguistics research, we will now turn our attention to aspects where there is more variance, as this helps to paint a more nuanced picture of the field. We will structure this discussion according to the discriminating criteria outlined by McEnery and Hardie (2012), by which corpus linguistics research can be divided and subcategorised. These criteria are as follows: mode of communication, corpus-based versus corpus-driven approaches, data collection regimes, annotated versus non-annotated corpora, total accountability versus data selection and multilingual versus monolingual corpora.

Mode of communication

Corpora can encode language that is produced in any mode, with existing corpora representing written, computer-mediated, spoken and gestural communication. Corpora representing computer-mediated communication (CMC) are the least challenging to construct, as the data are readily available in a machine-readable format, which means that the material collected usually requires minimal (if any) preparation to be rendered into a corpus (King 2009). An example of such a corpus is the Cambridge and Nottingham E-Language Corpus (CANELC), a 1-million-word corpus of digital communication in English, taken from discussion boards, blogs, tweets, emails and Short Message Services (SMS) (Knight

et al. 2014). Some researchers also make use of the World Wide Web as a mega-corpus, as explored later in this section.

Corpora containing written language are more time-consuming than corpora of CMC to construct and are more prone to error, particularly if the corpus compiler is required to manually type out the texts (Smith et al. 1998). However, the advent of optical character recognition (OCR) software, coupled with the increasing availability of written texts in digital formats, mean that compiling written corpora has never been quicker or easier. Issues pertaining to copyright can pose a challenge when constructing large corpora of written (and some online) texts, with researchers often required to seek permission or pay a fee when collecting texts that are protected by copyright (see Chapter 3 of McEnery and Hardie (2012) for a discussion of legal constraints on corpus building). An example of a written corpus is the BE06 (Baker 2009), which contains just over 1 million words of general written British English texts sampled from the press, general prose, learned writing and fiction published between 2003 and 2008.

Compared with written corpora, spoken corpora are infinitely more challenging to construct, as the processes of recording and transcribing spoken language can be time-consuming, laborious and expensive (Adolphs and Carter 2013: 8–21). An example of a spoken corpus is the Spoken British National Corpus 2014 (Spoken BNC 2014; Love et al. 2017) which contains 11 million words of everyday conversations taking place in Britain between 2012 and 2016.

Relatively recent advancements in visual recording technology have facilitated the construction of corpora containing paralinguistic communication, for example, those analysed by Knight (2011) to study gestural displays of active listenership and Johnston and Schembri (2013) to study sign language. These corpora arguably pose the greatest technical challenge to construct, as they present all the difficulties associated with compiling spoken corpora, but also require researchers to obtain visual data and, crucially, work out how to best represent speech and gesture simultaneously, in all their multimodal complexity (Knight 2011, Ch. 2). Researchers compiling corpora of spoken or gestural communication will likely have to obtain the informed consent of participants before collecting any data. That data will usually have to be anonymised, with personally identifying features removed before it is published, either in its entirety or as excerpts in published research. This poses a further challenge for the multimodal corpus builder. While written corpora can be anonymised quite straightforwardly – by removing references to names, places and other personally identifying information – the task of anonymising corpora containing speech or visual recordings of people is eminently more challenging. We again refer readers to Chapter 3 of McEnery and Hardie (2012) for a discussion of ethical practice in corpus building.

Because of these technical challenges, corpora representing written (and increasingly computer-mediated) language tend to be larger and greater in number than their spoken and gestural counterparts. Although discriminating between corpora on the basis of mode is useful for this discussion, it should be borne in mind that these distinctions (particularly between spoken and written communication) do not map easily onto every corpus, as many existing corpora contain language from more than one mode. Perhaps the most famous example of this is the aforementioned BNC, which contains 90 percent written and 10 percent spoken texts. The same combination of written and spoken texts, in a similar ratio, is to be included in an updated version of the BNC (BNC 2014), currently under construction at Lancaster University. Although the boundaries can therefore be fuzzy, categorising corpora according to mode can at least be linguistically meaningful, since different modes

have been observed to exhibit distinct lexical and grammatical characteristics (Carter and McCarthy 2017).

Corpus-based versus corpus-driven linguistics

Another criterion by which corpus linguistic studies are often described and categorised is the distinction between those which take corpus-based and corpus-driven approaches (Tognini-Bonelli 2001). According to this taxonomy, corpus-based studies are those which use corpus data to explore an existing (pre-corpus) theory or hypothesis to either validate, refute or refine it. This approach to the use of corpus data is underpinned by the view that corpus linguistics is a method. Corpus-driven studies, on the other hand, set about generating hypotheses about language on the basis of the corpus data alone. This approach regards the corpus as embodying its own theory of language (Tognini-Bonelli 2001: 84–85). This is a distinction that has endured in the field and can be observed in the description of a great deal of corpus linguistic research to this day. However, this taxonomy has also been challenged in recent years, for example by McEnery and Hardie (2012: 147–152), who reject the notion that the corpus itself has a theoretical status, and so with it the notion that a distinction can be drawn between corpus-based and corpus-driven linguistics (accordingly, they argue that all corpus linguistics studies are necessarily corpus-based).

Data collection regimes

Paramount to the validity and success of any corpus linguistic study is the construction or selection of a corpus that is suited to meeting the aims of the research being undertaken. With this consideration in mind, the issue of data collection becomes critical. There are two broad approaches to corpus data collection: the monitor corpus approach (Sinclair 1991: 24–26) and the balanced corpus or sample corpus approach (Biber 1993).

In the monitor approach, the corpus continually expands to include more and more data over time. An example of a monitor corpus is the Bank of English (BoE) (Hunston 2002). The BoE represents a mixture of general English and English language teaching materials which began to be collected during the 1980s and has been continually expanded ever since. At the time of writing, the corpus contains 450 million words. However, the largest monitor corpus available today is the World Wide Web (Gatto 2014). In addition to using standard search engines, such as Google, the Web can be explored as a corpus using dedicated interfaces, for example, WebCorp (Renouf 2003). However, the size and currency of the web as corpus (new texts are added to it constantly) must be offset against the fact that tools like WebCorp do not discriminate texts according to type, instead treating the Web as one homogeneous genre, which means that the data obtained are likely to require significant manual processing to sort into meaningful groups of texts.

In the balanced or sample approach, the corpus is constructed according to a specific sampling frame so that it represents language as it exists at a given point in time. Sampling frames are used to ensure that the corpus is balanced and representative in terms of the linguistic varieties or populations that it aims to represent. An example of this is the Brown family of corpora. Compiled during the 1960s, the Brown Corpus (Kučera and Francis 1967) contains over 1 million words of edited English prose printed in the United States in 1961. The Brown Corpus inspired the development of a number of other English language corpora, commonly referred to as the Brown family – so-called because they were all built

using the same sampling frame as the original Brown Corpus. This includes the Lancaster-Oslo/Bergen Corpus of standard written British English during the 1960s (LOB; Johansson et al. 1978), as well as the Freiburg-Brown corpus of American English (Frown; Hundt et al. 1999) and the Freiburg-LOB Corpus of British English (FLOB; Hundt et al. 1998), which, respectively, represent American and British written English during the 1990s. The benefit of building a corpus according to a sampling frame is that it can then be compared more systematically against other corpora designed using the same sampling frame. In this case, the Brown family of corpora provide the opportunity for studying change and international variation in contemporary written modern English.

Yet even the distinction between monitor and sampled corpora is not hard and fast. Some corpora incorporate elements of both approaches into their design, an example being the Corpus of Contemporary American English (COCA) (Davies 2009), which expands over time like a monitor corpus, but does so according to specific sampling criteria, with each addition containing the same breakdown of text varieties. Another type of corpus that does not readily fit the distinction between monitor and sampled/balanced corpora is opportunistic corpora. These corpora neither adhere to specific sampling frames nor address issues of skew by collecting an ever-increasing body of data, but instead seek simply to represent all the relevant data that were available to meet the research aims. Opportunistic approaches to corpus compilation are useful for cases where there is a limited amount of data available, such as when researching minority or near-extinct languages, for example the development of the EMILLE corpus (Baker et al. 2004) which comprises a series of monolingual sub-corpora representing fourteen South Asian languages, totalling 96 million words.

Annotated versus non-annotated corpora

Once they have been assembled, corpora can be encoded – or ‘annotated’ – with additional information relating to the texts or language(s) they contain (Leech 1997a). Annotation at the text level is likely to indicate the genre or mode of the texts, or demographic metadata about the authors/speakers, *inter alia* speaker sex, age and social class (Baker 2010). Meanwhile, annotation at the linguistic level usually indicates a word’s grammatical class (Leech 1997b) or semantic field (Wilson and Thomas 1997). Linguistic annotation is normally carried out automatically using so-called ‘taggers’, such as the USAS tagger integrated within *Wmatrix* (Rayson 2008). Annotation marks another way in which corpora and corpus linguistic research can be divided. Some corpus linguists see value in annotation as offering greater searchability, for example, in terms of the semantic and grammatical patterns in the corpus (Leech 1997a). However, others take a less favourable view of annotation, arguing that it imposes on the theoretical purity of the corpus and has the potential to be inconsistent or inaccurate (Sinclair 1992). In response to the latter criticism, it should be noted that it is possible for the researcher to manually check and correct erroneous tags assigned to a corpus by a computational tagger. However, we should also bear in mind that the larger the corpus we are dealing with, the more laborious and less practical this process of manual checking becomes.

Total accountability versus data selection

An advantage of corpus linguistics methods is that they allow us to take a more scientific approach to the study of language. For Leech (1992: 112), a core principle underlying the scientific approach to corpus linguistics is total accountability; that is, that the corpus

analyst/builder should not consciously gather data or focus on instances which support their argument (i.e. they should avoid confirmation bias). Using such data, hypotheses are, in principle, falsifiable – we do not simply look at evidence we know will confirm the hypothesis, we look at data collected without regard to ‘proving’ the hypothesis, hence the data should allow us to better explore if a linguistic hypothesis is falsifiable – an important feature of the scientific method (Popper [1935] 2006). Leech (1992) suggests that one way to satisfy Popper’s criterion of falsifiability is to test hypotheses and develop theories based on *all* relevant evidence emerging from the corpus in its entirety. Where there is too much evidence to do this (as is often the case with corpora that are vast in size), Leech suggests that we can nevertheless preserve the principle of total accountability by (1) avoiding conscious selection of data and (2) basing analyses on randomly randomly selected (i.e. non-biased) subsamples of the data.

However, attempts at total accountability in corpus linguistics can be criticised on the grounds that the corpus itself typically represents a necessarily finite subset of a much larger, non-finite entity (i.e. language). And this subset has itself resulted from a necessary selection and screening process. There are exceptions to this, for sometimes the amount of evidence required to test a hypothesis is finite (e.g. the published work of Charles Dickens), in which case it is possible to build a corpus that contains all qualifying/relevant texts and so is maximally representative of the type of language under investigation. However, in most cases claims of total accountability in corpus linguistics research must be moderated, for the analyst is usually only ever able to be totally accountable for the finite, screened sample of language contained in their corpus. Another challenge of using limited datasets is that the screened samples that corpora essentially represent might, by accident or design, offer a false or skewed representation of the wider body of language that they claim to represent. Methodological transparency (in terms of data collection and procedural choices) is therefore crucial in the reporting of a corpus linguistic study to ensure that that study can be *replicated* and the findings tested and verified by other researchers using other datasets.

It is important to acknowledge, however, that it is not the intention of all researchers using corpora to be totally accountable in terms of reporting the patterns in their data, or even to produce analyses that are replicable. A good example of this is research combining corpus linguistics methods with critical discourse analysis (CDA) (Fairclough 1995; see also Chapter 11). CDA seeks to examine the ways in which language and communication relate to issues of power and inequality, usually by providing detailed and contextualised accounts of small datasets. Since it is seldom practical to undertake such granular analyses on the large numbers of texts that tend to be contained within corpora, CDA researchers often either use corpora as a repository of interesting or relevant examples (Flowerdew 1997) or use quantitative and statistical measures to down-sample and focus on those features that are salient and relevant to their research aims (Baker et al. 2008). In such cases, the researcher(s) cannot claim total accountability for their data (a necessary compromise, perhaps, when combining these methods). However, by being transparent about their methodological choices and procedures, such studies can at least be replicated and their findings tested by others.

Multilingual versus monolingual corpora

The final discriminating factor we consider here is the distinction between multilingual and monolingual corpora. Most corpora are monolingual, containing texts that have been spoken, written or typed in just one language. However, some corpora are multilingual and

strive to represent more than one language. McEnery and Hardie (2012: 19) propose that multilingual corpora can be further subcategorised into three main types: (1) Those that contain source texts in one language plus translations into one or more other languages (e.g. the Crater Corpus (McEnery and Oakes 1995)); (2) Pairs or groups of monolingual corpora designed using the same sampling frame (e.g. the Brown family of corpora, discussed earlier); and (3) A combination of (1) and (2) (e.g. the aforementioned EMILLE corpus of South Asian languages (Baker et al. 2004)).

Like most of the discriminating criteria outlined in this section, the distinction between monolingual and multilingual corpora is not always clear-cut and may be a matter of degree rather than an absolute. Indeed, there are cases where corpora that are designed to be monolingual inadvertently contain words and speech in other languages. For example, while the bulk of the Early Modern English Books Online corpus (see McEnery and Baker 2016) is in English, the corpus also includes complete texts in Dutch, French, Italian, Latin and Welsh. On a smaller scale, the Spoken BNC 2014 (Love et al. 2017) contains numerous words and phrases from languages other than English, such as Bulgarian, Chinese and Polish. However, McEnery and Hardie (2012: 18) suggest corpora such as these should not be regarded as multilingual, but that the title be preserved for corpora that have been compiled with the conscious and deliberate aim of representing all the languages it contains.

The defining features of corpus linguistics set out in this section provide an overview of the various ways in which corpora and corpus linguistic studies vary and can be described and categorised. Despite the oppositional (even binary) nature of some of these features, the foregoing discussion does not mean to paint a picture of corpus linguistics as a discipline divided. Rather, this dialogue and debate, which has characterised much of the corpus linguistics literature in recent years, can be interpreted as reflecting the maturity of the field. Indeed, intra-disciplinary variance and disagreement of this sort is a luxury reserved only for established disciplines and has resulted from the relatively recent but growing acceptance and popularity of corpus linguistics within and beyond linguistics.

Current contributions and research

Having outlined some of its defining features, we now consider the current and potential contribution of corpus linguistics to the digital humanities, comparing it against other, more established approaches within the field. However, beginning with its contribution to linguistics more specifically, it is difficult to overstate the impact that corpus linguistics has had on the study of language. Indeed, it is no exaggeration to state, as Leech (2000: 677) does, that the availability of large datasets – and, more specifically, the tools with which to analyse them – has ‘revolutionised’ the ways in which we approach and now even conceptualise language. Corpora have been utilised by researchers working in a wide range of areas of linguistics, including but not limited to contrastive linguistics (Johansson 2007), discourse analysis (Baker 2006), CDA (Baker et al. 2008), language teaching/learning (Flowerdew 2012), semantics (Ensslin and Johnson 2006), pragmatics (Aijmer and Rühlemann 2014), sociolinguistics (Baker 2010), historical linguistics (Leech et al. 2009), theoretical linguistics (Xiao and McEnery 2004), literary linguistics (Mahlberg 2013) and psycholinguistics (Ellis and Simpson-Vlach 2009).

On paper at least, corpus linguistics and the digital humanities appear to be a good match: both are inextricably tied to digital technology, both use digital or digitised data and both use computational tools for analysis. Yet the place of corpus linguistics within the digital

humanities, and the contribution of corpus methods to digital humanities research, appears limited at present. As Jensen (2014: 115) points out,

[m]aking use of both digitized data in the form of the language corpus and computational methods of analysis involving concordancers and statistics software, corpus linguistics arguably has a place in the digital humanities. Still, it remains obscure and figures only sporadically in the literature on the digital humanities.

This conspicuous absence is not without cause. While there are certainly similarities between corpus linguistics and more established approaches within the digital humanities – approaches to studies in the humanities using a range of computational methods to automate or assist the analysis of an object of study – there are also significant differences.

The best way to demonstrate such differences, but also similarities, is to compare corpus linguistics against methods that are more established in the digital humanities. In this section we will look at an emerging method in the digital humanities known as *culturomics*. *Culturomics* represents a useful contrast to corpus linguistics as, unlike some other approaches in the digital humanities, it clearly emphasises the computational and automatic over more blended approaches to analysis. Spurred by a milestone paper by Michel et al. (2011), *culturomics* is the ‘application of high-throughput data collection and analysis to the study of human culture’ (p. 181). With the help of computational tools, it sets out to ‘observe cultural trends and subject them to quantitative investigation’ (p. 176). The authors of this article also argue that *culturomic* results provide a ‘new type of evidence in the humanities’ (p. 181), extending ‘the boundaries of scientific inquiry to a wide array of new phenomena’ (p. 176). McEnery and Baker (2016: 13–14) observe that *culturomics* has become something of an industry with researchers, typically not linguists but rather from the physical and other sciences, taking the *n*-grams available from the search engine provider Google and using statistical techniques to make (rather broad and sweeping) statements about culture through language.

To explore the similarities and differences between *culturomics* and corpus linguistics, we will compare the aforementioned study by Michel and colleagues against a corpus linguistics study, specifically that reported in Chapter 4 of Baker (2014). We selected the study by Michel et al. because, as previously stated, it has spurred interest in *culturomics* and provides what is, for us, a representative example of how the method is applied to study culture in language. We decided to compare this with the study by Baker because, likewise, it provides a representative example of how established corpus techniques can be used to study culture through language. Both studies account for written texts published within similar time periods: Michel et al.’s data covered the period 1800 to 2000, while Baker’s spanned 1810 to 2009. Finally, and most importantly, there is overlap in terms of the topic of Baker’s study and one of the cultural themes, or ‘trajectories’, explored by Michel and colleagues; that is, the frequency of gendered nouns (*man/men* and *woman/women*) in written texts over time.

Beginning with Michel et al., this research aimed to study cultural developments as they are reflected in language change over time. The analysis was based on a collection of just over 5 million digitised books (accessed through Google Books), which the authors claimed represented approximately 4 percent of all books published as of 2011. These data were a subset of the available total of 15 million digitised books. This subset was selected according to two criteria: (1) the accuracy of OCR and (2) the metadata available about the book. In these data, the analysis set out to identify ‘both linguistic changes, such as changes in

the lexicon and grammar, and cultural phenomena, such as how we remember people and events' (Michel et al. 2011: 176). The analysis was entirely frequency-based and compared the occurrences of a series of researcher-determined cultural keywords which corresponded to a wide range of fields, including historical epidemiology, the history of the American Civil War, comparative history, history of science, historical gastronomy and the history of religion. However, for the purpose of our comparison, we are most interested in results relating to gender. This part of Michel et al.'s analysis compared the frequency of the words *men* and *women* in the Google Books data over a 200-year period (1800–2000). They found that men were consistently mentioned more than women, but that the gap in frequencies narrowed over time, with mentions of *men* decreasing and mentions of *women* increasing, eventually to the point where mentions of women overtook those of men by the twenty-first century. On this basis, the authors conclude that, '[i]n the battle of the sexes, the “women” are gaining ground on the “men”' (Michel et al. 2011: 182).

Baker (2014) studied the frequency of gendered nouns in the Corpus of Historical American English (COHA) (Davies 2012), a corpus which consists of approximately 400 million words, taken from over 100,000 American English texts published between 1810 and 2009. The corpus was divided into four sections: fiction, magazines, newspapers and non-fiction, from sources which include Project Gutenberg, as well as scanned books and film and play scripts. The texts in the corpus had been tagged for part of speech and categorised according to year and decade of publication. Baker reported similar frequency patterns to those observed by Michel and colleagues. Examining the lemmas MAN and WOMAN, that is, all the words sharing the lexical root man (*man*, *man's*, *men*, *men's*) and woman (*woman*, *woman's*, *women*, *women's*), Baker found, like Michel et al., that male nouns were consistently more frequent than female nouns. He also found that the gap in frequency narrowed over time, reaching its narrowest point by the 2000s. However, mentions of men remained more frequent than women in his data. Therefore, generally speaking, the findings from both studies of the frequency of gendered nouns were similar, showing a decrease in the gap between mentions of women and men over time. The consistency of these findings is rendered more striking when we compare graphs which display this frequency trend produced by either study (Figure 20.1).

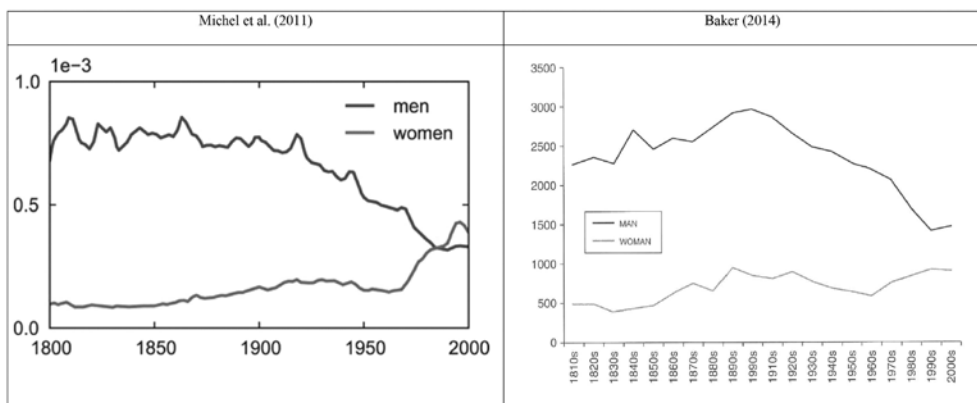


Figure 20.1 Frequency of gendered nouns in Michel et al. (2011) and Baker (2014)

With the help of more qualitative corpus techniques which enabled the analysis of the gendered nouns in their original textual contexts (introduced later in this chapter), Baker widened his search to consider the representation of men and women in the contexts of relationships and jobs. In terms of relationships, he found that, when mentioned in quick succession, MAN tended to precede WOMAN. He also found that women were more likely to be referred to in terms of their relationships to men (e.g. as *wives*) than men were in terms of their relationships to women (e.g. as *husbands*). In terms of job titles, gender-neutral terms (e.g. *chair* or *chairperson* versus *chairman*) were increasing in frequency. However, again examining those terms within their original textual surroundings, he noticed that those terms rarely related to women. Following this trend, when job titles were gendered (e.g. *chairman* as opposed to *chairperson*, *policeman* as opposed to *police officer*, etc.), they were 40 times more likely to incorporate male rather than female nouns, regardless of whether those jobs were actually fulfilled by men or women. Baker also found that jobs that were male-gendered tended to be positions of influence or leadership (e.g. *chairman*, *foreman*, *spokesman*), related to ensuring the safety of others (e.g. *policeman*, *fireman*) or required some form of transport to be controlled (e.g. *seaman*, *horseman*, *coachman*). On the other hand, women's jobs tended to involve service-related tasks, often relating to clothing and cleaning (e.g. *needlewoman*, *washwoman*, *charwoman*). Jobs that were exception-marked revealed implicit assumptions about the usual genders associated with them (e.g. *male nurse*, *female reporter*), while powerful jobs associated with women (e.g. *chairwoman*) didn't emerge until the 1980s. From his analysis, Baker concluded that there is an 'overwhelming male bias in American English, stretching back at least as far as the start of the nineteenth century' and that 'although the second half of the twentieth century has indicated a move towards equalisation, . . . this process appears to be ongoing and incomplete at the time of writing' (2014: 18).

Overall, then, there were some interesting similarities and many differences between the findings reported in the culturomics and corpus linguistic studies. The narrowing gap between the frequencies of the gendered terms was interpreted by both studies as reflecting a narrowing gender gap in society. However, Baker went on to provide a more nuanced (and ultimately less positive) picture of gender equality in society, revealing an 'overwhelming male bias' (2014: 18) in terms of how the gendered nouns in his data were actually used. There are several reasons why Baker was able to provide a more detailed account of his data, all of which mark significant points of methodological divergence between corpus linguistics and culturomics – as well as other methods in the digital humanities – which we will consider now.

The first difference concerns the collection and preparation of the data. Although both studies examined digitised data, there were significant differences in the size and structure of those datasets. Respecting size, the 5 million books analysed by Michel and colleagues is significantly larger than the 100,000 texts that made up the data that Baker analysed. However, size does not always equate with representativeness. One criticism that could be levelled at the Google Books data that Michel and colleagues analysed is that it represents one relatively narrow field of experience (books hosted by Google), and so, it might be argued, offers a rather small porthole through which to view and study 'culture'. However, it is difficult to accurately assess the representativeness of Michel et al.'s collection of digitised books when it is not possible to discriminate the texts it contains according to genre. The ability to discriminate the texts in a corpus in this way is important for linguistic research, given that genre has been shown to influence the frequency of certain words (Biber et al. 1999), as well as their immediate textual surroundings (or 'collocates', explored later in the chapter) (Gablasova et al. 2017). Where corpora are

principally structured bodies of digitised text, the data that are collected and studied as part of much digital humanities research – including culturomics studies – tend to represent unstructured bodies of “raw” text which have undergone little, if any, processing, and so are considerably less principled in their design (Jensen 2014). Treating all the texts in the data as one genre, as the culturomics approach forces us to do, meant that Michel and colleagues were not able to account for the extent to which the frequency patterns they observed were influenced by the genre of the books. Therefore, although Baker’s corpus is smaller, its principled design and structure meant that he was better aware of what his corpus did (and, just as importantly, did not) represent, and he could thus have greater confidence about his analytical claims.

The second major difference between the culturomics and corpus linguistics approaches was that the former relied on the human researchers’ *a priori* knowledge of which words and word combinations to search for (in this case, just *men* and *women*). Thus, the analysis could only account for those words that the researchers intuitively regarded as culturally significant and so actively searched for. Although Baker’s search terms were also decided *a priori*, corpus linguistic methods allowed him to go beyond his original search terms to explore patterns that were prevalent in his data but of which he was not necessarily cognizant prior to his analysis. Accordingly, Baker’s analysis was able to account for a wider range of gendered nouns, some of which – particularly relationship and occupational nouns – revealed significant gender ideologies that did not emerge from analysis of the lemmas MAN and WOMAN alone. Although it may not have been Michel et al.’s aim to look at any words other than *men* and *women* in their analysis, the lack of opportunity to go beyond researcher-determined, largely intuition-driven search terms nevertheless presents some issues with the culturomics approach. Relying solely on human intuition when deciding on analytical search terms means that you will only be able to explore patterns that you actively search for, based either on intuition or *a priori* knowledge of the language or texts under study. In other words, unless you search for a pattern, you may not see it. Yet unless you are aware of the pattern, you will be unlikely to search for it in the first place (McEnery and Baker 2016: 12). Adopting this approach, we also run the risk of assuming that any finite number of words is sufficiently representative of a particular cultural phenomenon that we wish to investigate. This concern becomes particularly pressing when studying historical data, since words’ meanings can change over time and existing words can fall out of use and be replaced by new ones.

A third difference between the culturomics and corpus linguistic approaches concerns the way that frequency patterns were interpreted. Although frequency can provide information about which words are used frequently in a language, there is no perfect relationship between how frequent a word is and how salient that word (or the concept it denotes) is within a society. For example, in a 2008 study, Baker showed that the high frequency of words denoting homosexuality in general language did not reflect the fact that the concept was regarded favourably within society but, on the contrary, that heterosexuality was assumed to be the norm and so cases of homosexuality became linguistically and culturally marked (Baker 2008: 146–148). Frequency information should therefore not be taken at face value, but instead scrutinised and interpreted by the human analyst using their knowledge of society and supported with analysis of how words of interest are actually used in the texts in the data. It is because Baker (2014) did this that he was able to provide a more nuanced explanation of the frequency patterns he observed – specifically that the narrowing in frequency of gendered nouns did not necessarily correspond with a surge in gender equality.

Fourth, and finally, a major difference between the approaches compared in this section is that culturomics does not allow users to examine how search terms are actually used by analysing them within their wider textual surroundings in the way that corpus linguistics methods do. Without access to the texts contained in the data, or even knowing the genres of those texts, it becomes extremely difficult to reliably interpret the meaning of any word or phrase of interest, not least with regard to the cultural trends that that word is interpreted as reflecting. Such a view of the data is not provided by the decontextualised view of words that is made available by Google in the form of n-grams, which reveal frequency information only and on which Michel and colleagues' analysis was based. Although corpus linguistics studies typically begin with the use of decontextualised examples (usually in the form of word lists), other weapons in the corpus linguist's arsenal (to be introduced in the next section) can provide a more contextualised view of the data, allowing a more sophisticated analysis to be undertaken and more nuanced (and ultimately, more accurate) conclusions to be drawn. For example, Baker (2014) reported an increase in the use in gender-neutral job titles in his data, one such word being *chairperson*. At face value, this trend might be taken as evidence of greater gender equality in terms of the people who were in those traditionally male-dominated roles in society. However, when he examined the occurrences of the word *chairperson* in context, Baker found that the chairperson being referred to could be identified as female in just 2 percent of cases in his sample, with the term more likely to refer to men, as indicated by the accompanying gendered pronouns *he* and *his*. In other words, although the job titles had become more gender-neutral, in reality the roles they referred to still tended to be dominated by men.

As our comparison between these two studies has hopefully showed, we would argue that there is limited value in assembling, as the culturomics approach would, a vast but unstructured mass of text for examining culture. This value is diminished further when we are forced to access that data using what are, in comparison to corpus linguistics methods, relatively primitive techniques which offer information about the frequency of the words contained in the data but deprive the researcher of the opportunity to look past the numbers and view and scrutinise patterns of language in-use. Although Baker and Michel et al. reported similar overall frequency patterns for gendered nouns, with the help of corpus linguistic methods Baker could inform his interpretation of those frequency trends with insights into how those words were actually used in the data, providing a more explanatory and ultimately more nuanced account. This is where corpus linguistics methods can be useful for the (digital) humanities; they can feed the appetite for quantitative methods and 'big data' which underpins approaches like culturomics, but triangulate these insights with a level of granularity that is associated with methods situated at the qualitative end of the spectrum, like close-reading.

Sample analysis

In this section of the chapter we will introduce a series of established corpus linguistic techniques, providing worked demonstrations of their use. For this analysis, we will draw on the NHS Comments Corpus (NHSCC) (Brookes and Baker 2017; Brookes and McEnery 2017); a corpus of patient feedback comprising all comments submitted to the NHS Choices online service over a period of 2.5 years (March 2013 to September 2015). In its entirety, the NHSCC is approximately 29 million words in size and contains 228,113 comments relating to a wide range of healthcare organisations, including acute trusts, care organisations, care providers, clinical commissioning groups, clinics, dentists, general practitioner

(GP) practices, hospitals, mental health trusts, opticians and pharmacies. For the purpose of this chapter, we will focus on a subset of the NHSCC, specifically comments relating to GP practices. This is the largest sub-section of the data and comprises 111,318 comments and 14,093,437 words. This analysis will demonstrate how corpus techniques can be used in an exploratory way. Although we therefore did not approach these data with any specific research questions in mind, we were broadly concerned with identifying and examining frequent themes of patient feedback in our data. As pointed out earlier, there is no standard approach or set of procedures in corpus linguistic methodology. However, our sample analysis will introduce and make use of four well-established techniques in corpus linguistics: frequency, keywords, collocation and concordance, all of which were accessed through the *#LancsBox* (Brezina et al. 2015) software package. A full analysis would employ these techniques (and more) to examine a wide range of concerns in this feedback data (see Baker, Brookes and Evans 2019). However, in this chapter we will demonstrate how these techniques can be used in an exploratory fashion to pinpoint frequent themes in large collections of textual data which can then be explored in more granular detail.

Frequency

Frequency provides us with a list of all the words that are present in our corpus and tells us how often each one occurs. Frequency information can provide a good starting point for corpus analysis, as it affords a rapid overview of the linguistic landscape of the data, including the most common linguistic items. Frequency information can be generated to account for individual words or recurrent sequences of two or more words. Many software packages also afford the possibility to obtain lemmatised frequency lists (introduced earlier). If the corpus is annotated, we can also generate a list of tags reflecting the most common grammatical or semantic types in the data. However, we will focus here on the frequencies of individual words occurring in our untagged corpus of GP comments. Table 20.1 displays the ten most frequent words in the corpus (an arbitrary cut-off).

As we might expect – given their high density in general language (Leech et al. 2009) – the most frequent items in the corpus tend to be grammatical words (e.g. *the*, *to*, *i*). Studying the frequencies of such words can be useful for learning about the genre and register features of the texts in the corpus (Biber and Conrad 2009). In terms of our corpus of GP comments, for example, we might interpret the high frequency of the first-person singular

Table 20.1 Top ten words in the GP comments, ranked by frequency

Rank	Word	Freq.	Freq. per million words
1	the	589,573	41,843
2	to	488,810	34,692
3	i	475,528	33,749
4	and	423,909	30,086
5	a	284,278	20,176
6	have	196,513	13,947
7	of	158,164	11,225
8	for	154,744	10,983
9	is	151,490	10,752
10	in	150,720	10,697

pronoun *i*, which is the third most frequent word in the corpus, as providing evidence that the people submitting feedback tend to do so from their own perspectives, and so presumably about their own experiences of healthcare, rather than those of friends or relatives. However, grammatical items, such as the ones dominating our frequency list, reveal comparatively less about the content of the texts in the corpus. For this, it is much more fruitful to analyse content or ‘lexical’ items (nouns, verbs, adjectives and lexical adverbs) (Baker 2006: 54).

Frequency information can be used to access lexical items, but doing so usually requires the analyst to filter their results and cast their gaze further down the frequency list to extract such words manually. For example, the most frequent lexical item in our GP comments is *surgery*, which is ranked nineteenth in the frequency list and occurs 111,555 times in the corpus. To learn about the content of the patients’ comments, we might continue in this vein, running our eyes further down the list and extracting lexical items and disregarding grammatical words. Such an approach has been used to good effect in corpus research, particularly that which uses frequency as a starting point for more qualitative discourse analysis (Baker 2006). However, corpus linguists are generally reticent to *prima facie* disregard any word or pattern in the corpus, particularly if that word occurs with a high frequency. How frequency information is used will rightly depend on the nature and specific aims of each study. However, a more statistically robust way of identifying words which reveal the content of the texts in our corpus is to use the keywords procedure.

Keywords

Keywords are words which occur with either a significantly higher frequency (positive keywords) or significantly lower frequency (negative keywords) in the corpus we are analysing when it is compared against another corpus (Scott 1997). The corpus that we compare our analysis corpus against is referred to as the reference corpus. The reference corpus typically represents a norm or ‘benchmark’ for the type of language under investigation. Words are deemed to be keywords by the computer based on statistical comparisons of the word frequency lists for each corpus. The choice of reference corpus is important here, as it shapes the keyword output. When selecting a reference corpus, we usually want one that is similar in size to, or larger than, the analysis corpus. Ideally, the reference corpus should also represent language belonging to the same genre as the texts in our analysis corpus, so that our keywords will flag up what is lexically distinctive about the texts in our corpus compared to others of a similar type.

For our analysis we compared our corpus of GP comments against the BE06 (Baker 2009), introduced earlier, which contains just over 1 million words of modern-day general written British English. Once we have selected our reference corpus, we also have to choose a statistic to measure the ‘keyness’ of each word in the corpus and decide whether or not we want to impose a minimum frequency threshold for candidate keywords. For our analysis, we generated keywords using log ratio (McEnery and Baker 2016: 23), which combines a test of statistical significance (log likelihood (Dunning 1993)) with a measure of effect size, which quantifies the strength of the difference between the observed frequencies, independent of the sample size. This statistic will produce as keywords those words whose relative (or ‘normalised’) frequency is significantly higher (beyond the 99.9 percent threshold level) in the corpus we are analysing compared to the reference corpus. Each statistically significant keyword is then assigned a log ratio score based on the size of the observed difference in relative frequencies, with bigger differences producing higher scores. This measure

Table 20.2 Top ten keywords in the GP comments, ranked by log ratio

Rank	Keyword	GP comments		BE06		Log ratio
		Freq.	Freq. per million words	Freq.	Freq. per million words	
1	surgery	111,555	7,915	24	21	8.56
2	rude	21,114	1,498	6	5	8.16
3	gp	58,880	4,178	21	18	7.83
4	receptionist	25,678	1,822	11	10	7.57
5	appointments	35,769	2,538	16	14	7.51
6	appointment	99,519	7,061	52	46	7.28
7	helpful	25,331	1,797	18	3	6.84
8	8am	3,709	263	3	3	6.65
9	unhelpful	5,374	381	5	4	6.45
10	doctors	57,061	4,049	55	48	6.40

therefore uses log-likelihood as a cut-off to ensure that keywords are significant, but has the advantage that it also allows us to rank those keywords according to how unusually high, or ‘marked’, their frequency in the corpus is. We also imposed a minimum frequency threshold of 3, meaning that a word had to occur at least three times in both corpora to qualify as a candidate keyword. The top ten keywords, ranked by log ratio score, are displayed in Table 20.2.

Before commenting on this table, it is useful to first clarify how these words have become keywords. Although frequency is a driving factor in identifying keywords, high frequency alone is not enough for a word to be judged as ‘key’. What is most important in determining whether a word is a keyword is the relative frequencies of the word between the two corpora being compared. For example, the most frequent noun in this list, *surgery*, occurs 7,915 times per million words in our GP comments. This is very much greater than its normalised frequency of 21 times per million words in the BE06. Therefore, the word *surgery* is a keyword in the GP comments when its frequency is considered not simply on its own, but against some norm of usage.

As Table 20.2 shows, the keyword measure has reliably filtered out the high-frequency grammatical items which dominated our frequency list, instead bringing to the fore those words which are most characteristic of our corpus and more revealing in terms of the content, or ‘aboutness’ (Scott and Tribble 2006: 59–60), of the comments it contains. One way of approaching these words is to group them into semantic categories reflecting the most characteristic themes in the corpus. In this vein, we might interpret these keywords as reflecting a focus on sites of healthcare (*surgery*); the tendency for people to use evaluative language in their comments (*rude*, *helpful*, *unhelpful*); a focus on social actors and specifically staff members (*gp*, *receptionist*, *doctors*); and a focus on appointments (*appointments*, *appointment*, *8am*). This list of keywords thus provides a rapid and replicable overview of key themes in the corpus, as these are reflected in its most characteristic words. However, looking at words in isolation, as in a frequency or keyword list, is limiting, as it provides a decontextualised view of the corpus and reveals nothing about how those words are actually used in the texts it contains. In this case, although the keywords reveal key areas of patient interest/concern, the keyword list alone cannot tell us why patients addressed these themes in their feedback, nor what they had to say about them. To gain such an understanding, we

need to take a more contextualised view of the keywords of interest and examine their use *in situ*. We can begin to do this using the collocation technique.

Collocation

Collocation, as conceptualised in contemporary corpus linguistics, is a linguistic device whereby words, by associating strongly with one another, become bearers of meaning by virtue of co-occurrence. Collocation is typically judged to exist by the use of a word association measure that tells us how often two or more words occur alongside one another and whether this association is notable as a sizeable effect in our data; that is, that the words have a measurably strong preference to occur together as opposed to being randomly associated. Following Firth's (1957: 6) dictum that 'you shall know a word by the company it keeps', corpus linguists have long sought to learn about words' meanings and patterns of use by examining those words in terms of the words with which they co-occur, or 'collocate'. Analysing the collocates of a keyword or other word of interest can therefore provide insight into the textual context surrounding that word in the corpus. For example, in a previous analysis of the full version of this corpus, Brookes and Baker (2017) examined the frequent collocates of words expressing positive and negative evaluation (for example, *good* and *bad*) to investigate what patients tended to evaluate positively and negatively in their comments.

To demonstrate how a collocation analysis might proceed, we will analyse the collocates surrounding the keyword *rude*. As a reminder, *rude* was the second-ranked keyword, with a log ratio score of 8.16. It occurred 21,114 times across the GP comments, or 1,498 times for every 1 million words in the corpus. Like the generation of keywords, deriving collocates requires the researcher to make a series of procedural decisions, for example, pertaining to the span, method of calculation and use of a minimum frequency threshold – all of which present variables that will ultimately shape the amount and type of words that are flagged up as collocates by the computer. First, we have to decide on the collocation span; that is, the number of words to the left and/or right of the search word within which we want to search for candidate collocates. Tighter spans will produce a smaller number of collocates which occur within closer proximity to the search word, whereas wider spans will produce a higher number of collocates, some of which might not occur in such close proximity to the search word. We searched for collocates of *rude* using a window of five words to the left and right of the search word (otherwise expressed as L5 > R5). This window is fairly standard in corpus research, as it provides a 'good balance between identifying words that actually do have a relationship with each other (longer spans can throw up unrelated cases) and [gives] enough words to analyse (shorter spans result in fewer collocates)' (Baker et al. 2013: 36).

We then have to decide how we will rank, score or 'cut-off' the candidate collocates. We can do this either by ranking the collocates according to frequency of co-occurrence, or by using a statistical measure. Corpus linguistic software packages offer a range of statistical measures for determining the exclusivity of a collocational pairing (i.e. that two words do not co-occur frequently simply as a result of them both being frequent in the corpus). Popular measures include mutual information (MI), MI3, Z-score, log-likelihood, log-log and log-ratio. Again, the choice of statistic will shape the collocates produced (for a discussion see: Gablasova et al. 2017). We derived collocates of *rude* using the MI statistic. MI determines collocation strength by comparing the observed frequency of each collocational pairing against what would be 'expected' based on the relative frequency of each word and the overall size of the corpus. The difference between the observed and expected frequency of co-occurrence is then converted into a score (MI score) which indicates the

Table 20.3 Collocates of 'rude', ranked by MI

Rank	Collocate	Collocation frequency	MI
1	unhelpful	2,354	8.07
2	abrupt	678	7.97
3	unprofessional	538	7.25
4	receptionists	2,836	6.60
5	extremely	1,495	6.52
6	receptionist	2,909	6.13

strength of collocation, with higher scores assigned to stronger collocations. We restricted our search to those collocates which were assigned a minimum MI score of 6 (Durrant and Doherty 2010).

Finally, we have to decide whether we want to impose a minimum frequency threshold. A lower threshold will produce larger amounts of collocates which can occur alongside the search word sparingly, whereas higher thresholds will produce a smaller number of more selective collocates. Most software packages operate with default thresholds of between 3 and 5; however, this can be adjusted by the user. The choice of threshold will largely depend on the frequency of the search word (higher thresholds can be used with more frequent words), as well as how selective the analyst wants to be. Since the frequency of our search word (*rude*) was high (21,114 occurrences), we could afford to be more selective and so decided to focus on just a small number of collocates for this analysis. Therefore, we used a particularly high threshold of 500.

Using these criteria, the computer derived six collocates of *rude*. These are displayed in Table 20.3.

The words in Table 20.3, though few in number, all collocate strongly with *rude* and provide some interesting insights into how people talked about rudeness in their comments. Evaluative words like *unhelpful*, *abrupt* and *unprofessional* might be used in a (near-) synonymous sense (Xiao and McEnery 2006) and could shed light on the reasons why staff are evaluated as rude (e.g. *unhelpful* could index an inability to meet patient requests, while *abrupt* might relate to communication skills). The intensifier *extremely* suggests a tendency for people to grade levels of rudeness in their comments, and this collocate presents the opportunity to explore if there are any differences between what is evaluated as *rude* and *extremely rude*, for example. Finally, the words *receptionists* and *receptionist* might suggest that staff working in reception tend to be talked about in relation to rudeness more often than other staff groups. However, to test and substantiate this hypothesis we would need to examine *how* these words are actually used in relation to each other – a perspective on the data that is not afforded by collocation alone but with the help of other corpus techniques (particularly concordance, introduced later).

Corpus linguistic software programmes are effective for rapidly highlighting the relationships between words in a corpus, measuring the distance between those words when they do co-occur, the frequency and exclusivity of these relationships and reliably indicating whether or not these relationships are statistically significant. Yet a limitation of most software packages, and so the research that uses them, is that they only consider collocation in terms of two words at a time. However, more recently, corpus linguists have endeavoured to explore the relationships that exist between more than just two words (Baker 2005,

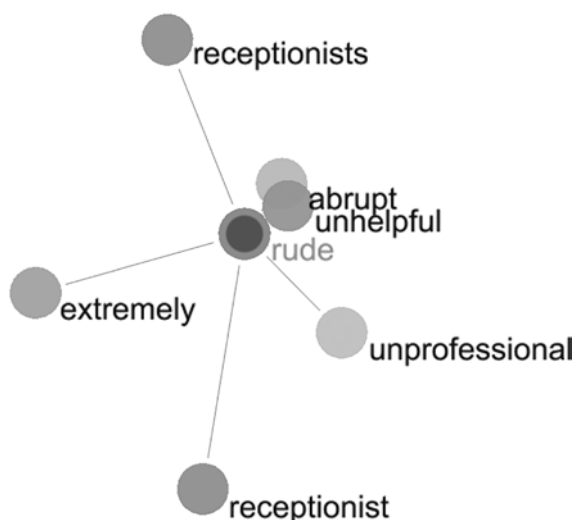


Figure 20.2 First-order collocates of 'rude'

2014; McEnery 2005, 2006) and to account for what are known as collocational networks. *#LancsBox* allows us to do this using GraphColl, a technique which shows collocational networks consisting of more than two words and visualises these networks in the form of graphs. Building on our analysis of collocates to this point, and retaining the parameters used to derive the collocates displayed in Table 20.3, we entered the word *rude* into GraphColl. The collocational network this generated is reproduced in Figure 20.2. We will refer to these as 'first-order' collocates; that is, the immediate collocates of a word (Brezina et al. 2015: 151). By extension, the collocates of *receptionists*, other than *rude*, if included in a graph such as this would be 'second-order' collocates of *rude*.

The length of each line in this diagram represents the MI score assigned by the computer to the respective collocational pairings, with shorter lines showing higher MI scores and so stronger collocational relationships. Therefore, we can see by comparing the lengths of the lines that *unhelpful* has a stronger relationship to *rude* than *unprofessional* does, as it is located closer to the search word in the centre of the graph. As well as visualising collocational networks, an advantage of *#LancsBox* is that it allows us to expand the collocational network to include second-order collocates (as noted, the collocates of the first-order collocates in the existing network). We expanded the network by introducing the collocates of the word *receptionists*, which is the strongest staff word collocate of *rude*. The expanded network is displayed in Figure 20.3.

Figure 20.3 displays a more complex network which helps to reveal a more nuanced picture of how receptionists are evaluated in the comments. We can see that, as well as collocating with *rude*, *unhelpful* is also a strong collocate of *receptionists*. The collocation of *receptionists* with *nurses* reflects the tendency for people to group these staff types together in their comments, and might indicate a gendered pattern of evaluation, since these are roles that are occupied mainly by women in the UK healthcare context. Finally, the collocation of

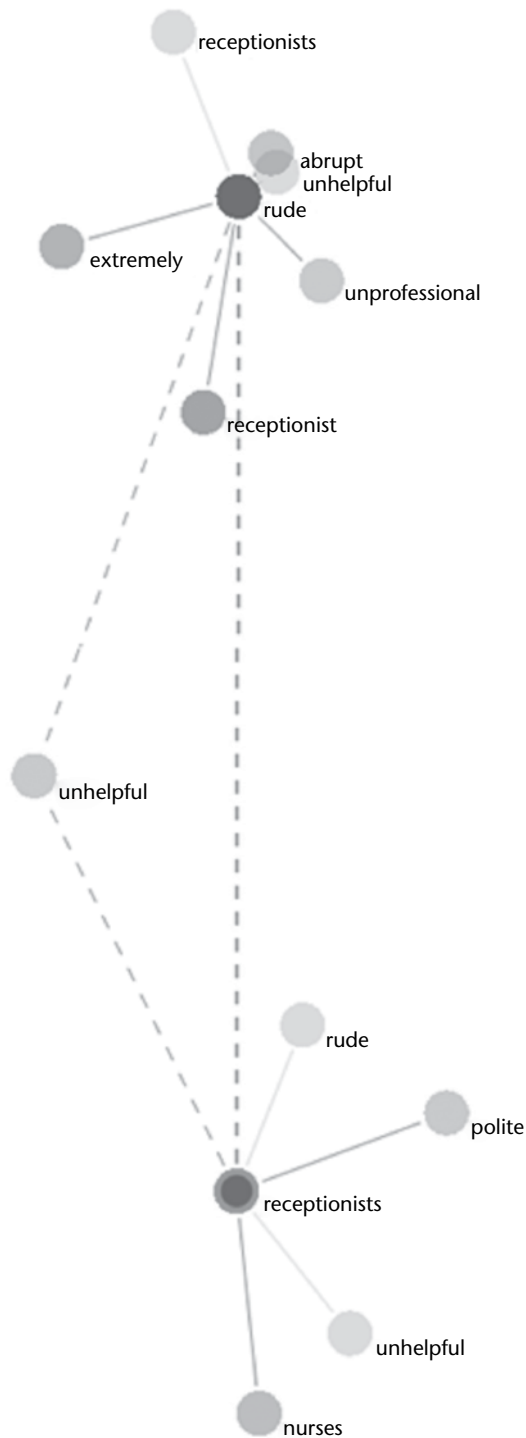


Figure 20.3 Second-order collocates when the search word is 'receptionists'

receptionists with *polite* provides evidence that receptionists are also frequently positively evaluated for their interpersonal skills, and so helps to paint a more balanced picture of feedback about this staff group, and one which would not have emerged had we not expanded the network to consider second-order collocates. In a full analysis, we could expand the network further, by introducing third-order, then fourth-order collocates and so on, each time building an increasingly complex collocational network (see, for example, Baker 2016).

Techniques for measuring collocation and generating collocational networks provide the opportunity to learn about words' meanings and patterns of use by inspecting their relationships to other words in the corpus. Such techniques therefore take us beyond the solitary linguistic item displayed in frequency and keyword lists and offer a logical 'next step' following these inductive analytical procedures. However, even with collocation we are still dealing with relatively de-contextualised samples of data, for even the most complex collocational networks provide only a narrow window into the often complex and nuanced relationships between words.

Concordance

The final method we wish to introduce here, concordancing, provides a way of viewing the corpus data that allows us to examine every occurrence of a word or phrase in context and thus quickly scan for patterns of use. Because it offers a more contextualised view of the data, this perspective is often referred to as a key word in context (or 'KWIC') view. Building on our analysis of collocational networks in the previous section, Figure 20.4 shows a random sample of concordance lines where *rude* collocates with *receptionists*.

With the search word running down the centre of the computer screen and a few words of context displayed to the left and right, the concordance output can be very useful for spotting patterns that might be less obvious during more linear, left-to-right readings of the data. The concordance lines can be displayed in order of occurrence, in random order or alphabetically according to the words surrounding the search word (rendering recurrent patterns more observable). For a more contextualised view of the data, the analyst can access each original text (in this case, comment) in its entirety, simply by clicking the highlighted search word in the centre of each concordance line. Concordancing is essentially therefore a means to adopt a different perspective on the data, providing the opportunity for human analysts to

Index	File	Left	Node	Right
9	gppractices-1	nearly impossible and the receptionists can be	rude	to you when there are no appointments
20	gppractices-1		Rude	Receptionists are very rude. Very unfriendly! Nev
21	gppractices-1	Rude Receptionists are very	rude	Very unfriendly! Never again.
23	gppractices-1		Rude,	uncating and unhelpful receptionists Treated like
34	gppractices-1	in advance- go figure the receptionists are	rude(	although not all) In my opinion they
54	gppractices-1		Rude	receptionists and an extremely rude Triage Nurs
55	gppractices-1	Rude receptionists and an extremely	rude	Triage Nurse The automated telephone system
62	gppractices-1		Rude	Receptionists Hello All, I just want to
64	gppractices-1	U cant ever get an appointment, v	rude	receptionists, avoid
75	gppractices-1	problem here a certain receptionists, they are	rude	un helpfull. Half the time they will
77	gppractices-1	Centre after 22 years due to very	rude,	grumpy receptionists, I was beginning to think
80	gppractices-1	surgery as I found the receptionists quite	rude	and one of the nurses fairly blunt
93	gppractices-1	called front of house receptionists have been	rude,	slapping there gums with chewing gum and
105	gppractices-1	other patients being abrupt and sometimes downright	rude	to the receptionists and yes, I appreciate
108	gppractices-1	you do get through the receptionists are	rude	and generally unhelpful. The surgery doesn't see
120	gppractices-1	later which is unacceptable! 4. Receptionists are	rude	and do not care about you. 5.

Figure 20.4 Concordance lines for collocation of 'rude' with 'receptionists'

undertake closer, more qualitative analyses of words or phrases in the corpus and develop theory-informed observations based on extended patterns of use.

To better understand why receptionists were evaluated as rude, we examined a sample of concordance lines in which the word *rude* collocated with *receptionists* (5L > 5R). The analysis of concordance lines can be approached in a number of ways. For this analysis, we followed Sinclair's (2003) recommendation of analysing a random sample of 30 concordance lines, recording observable patterns, then repeating the process until new patterns ceased to emerge. Although at face value the tendency for the word *receptionists* to collocate with *rude* might suggest that staff working in this area have a habit of interacting with patients in a way that is (at least perceived to be) rude, our qualitative analysis of the comments revealed that evaluations of reception staff as *rude* rarely related to their interpersonal skills, but tended to have more to do with, amongst other things, problems with the availability of appointments. For example, one comment reads:

We need more GP 's at this surgery. Appointments I have never had a problem with the medical team, they have been very thorough and have been very helpful. However, getting an appointment is nearly impossible and the **receptionists can be rude** to you when there are no appointments available. On one occasion I needed an emergency appointment, and so I waited in the waiting room for 2 hours, only to be told that the receptionist had not registered that I was there.

Though related to interpersonal and communication skills in terms of style of message delivery, our analysis of the concordance lines revealed that evaluations of receptionists as *rude* are therefore likely to have less to do with their interpersonal skills and more to do with their roles as medical gatekeepers. Receptionists act as a 'key buffer between practice and patients' (Litchfield et al. 2016: 3) which means that they are often the unfortunate 'bearers of bad news' when it comes to informing patients about the limited availability of appointments, as well as delays and cancellations. Thus, it is receptionists who frequently bear the brunt of patients' frustrations when issues around appointments arise. Yet it is important to bear in mind that booking and managing appointments is a complex task that requires receptionists to negotiate numerous factors, including balancing patients' needs and expectations with the actual availability of appointments. The tendency for receptionists to be described as rude might also be explained by the ways that patients' expectations of staff differ according to status and gender. It is possible that patients might be more lenient in their evaluation of 'rude' doctors than receptionists because they perceive the former to be more qualified and as capable of directly remedying their ailments. Similarly, since the role of medical receptionist is mostly taken up by women, the tendency for receptionists to be evaluated in terms of their interpersonal skills (that is, as *rude*) more often than other staff groups might reflect ingrained societal expectations that women will be more caring and polite than men (Mills 2003).

The previous section has demonstrated how established corpus linguistic techniques can be used to carry out in-depth quantitative and qualitative examinations of large collections of digitised language data. Using as our data an extensive body of online patient feedback, we combined these techniques in a cyclical approach that involved (1) using the frequency and keywords functions to identify general themes in the data and (2) using collocation to develop a sense of how the theme of rudeness was discussed in the patients' comments before (3) closely reading concordance lines where *rude* occurred with *receptionists* to confirm or revise our hypotheses and develop a more refined understanding of how patients discussed

this issue in their comments. During this final stage, to account for the patterns we observed and explain why receptionists were described as rude, it was fruitful for our analysis to go beyond the corpus itself and draw on our knowledge of receptionists' expanding roles, as well as how patients' expectations might be influenced by staff members' professional status as well as their gender. Where the computer had flagged up the saliency of the concept of rudeness, and even helped us to establish its relationship to receptionists, it was up to us, the human analysts, to dig deeper – by engaging with the comments qualitatively – and wider – by drawing on corpus-external knowledge and sources – to interpret and explain the patterns of evaluation. Only by doing this were we able to provide an account of the feedback which was, though necessarily limited, eminently more nuanced and balanced than could have been achieved using quantitative and computational methods alone.

Future directions

In this chapter we have introduced corpus linguistics and aimed to demonstrate how corpus methods can be useful to digital humanities researchers. McEnery and Hardie observe that '[t]here are few areas of linguistics where there is no possible role for corpus methods . . . and there are increasingly many where corpus methods have become central' (2012: 236). As we hope is clear from this chapter, the same can be said for the digital humanities. Corpora and corpus techniques can be useful for researchers from any disciplinary background who want to carry out more objective and statistically robust analyses, and who see value in basing their findings on larger and more representative bodies of data. Although the examples we have provided focus largely – perhaps unsurprisingly – on studies from linguistics, the reach of corpus linguistics extends beyond its own disciplinary confines, with corpus approaches notably applied to research in (digital) humanities subjects including but not limited to anthropology (Nolte et al. 2018), geography (Gregory et al. 2015), history (McEnery and Baker 2016), literature (Biber 2011) and translation studies (Laviosa 2002).

Due to limitations of space, we had to limit our introduction to just four of the most established methods in corpus linguistics (frequency, keywords, collocation and concordance), the use of which we demonstrated on only one type of data (online patient feedback). However, it should be noted that there is a wider and growing body of available techniques which can be combined in a multitude of ways and applied, in principle, to any type of textual data. Whichever method or combination of techniques we choose to adopt, critical to the corpus approach is the interplay between, on the one hand, computational and statistical measures and, on the other, human-led, theory-sensitive readings of the data. In corpus linguistics the computer does not 'do' the analysis for us. Corpus methods can help us to adopt novel perspectives on larger amounts of data to arrive at more interesting and statistically substantiated conclusions. However, it is the role of human analysts to decide which procedures to use and how to use them, and then to interpret the significance of the corpus output by engaging in close reading and drawing on their wider knowledge of the social contexts in which the texts were produced. Analysts should also take care when making grand claims about objectivity. It is true that computer programmes do not make errors and are not subject to ideological and cognitive biases in the same way as human minds. However, both the designers and users of these programmes *are* subject to such biases. It is therefore important for analysts to avoid uncritical overreliance on corpus techniques and be self-reflexive about the influence that their own methodological choices and biases are likely to have had on their findings.

To no small extent, it is the limitations of corpus linguistics, along with developments in computer technology, that have stimulated the most significant advances in corpus linguistics to date. One area where corpus linguistics arguably trails other methods of linguistic analysis is in its ability to account for non-linguistic, but particularly visual, modes of communication. Multimodal corpus linguistics has come a long way in a relatively short space of time. However, techniques for studying imagery, fonts, colours and so forth have yet to come as far as those which track modes like tone and gesture. Another direction for future research is in the development and refinement of techniques for data visualisation. The development of tools like *#LancsBox*, demonstrated in the analysis here, have marked significant strides in the capacity of corpus linguistics to capture the nuances and messiness of complex lexical relationships. This is an exciting juncture and we would expect to see further development of tools to this end. Another emerging research agenda for corpus linguistics is its triangulation with other approaches, both within and outside linguistics. To offer a recent example, McEnery and Baker's (2016) corpus-assisted study of the representation of prostitution during the seventeenth century provides a demonstration of the benefits that can arise when corpus linguists and historians work together. For more on the combination of corpus linguistics and history, including the approaches that historians bring to this synergy, see Chapters 1 and 7 of McEnery and Baker (2016). Whatever your discipline, corpus linguistics offers a powerful toolkit to researchers trying to get to grips with the increasing availability of vast bodies of textual data. If insights can be gained by studying text and there is value to be had in taking more objective approaches to larger and more representative bodies of data, then corpus linguistics will be able to add new dimensions and fresh perspectives to your research.

Further reading

- 1 Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.
Each chapter of this book demonstrates how the main techniques of corpus linguistics can be used to analyse discourse by means of a series of case studies.
- 2 McEnery, T., Xiao, R. and Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.
This book provides a hands-on introduction to corpus search software and includes a collection of important and interesting corpus linguistics research papers.
- 3 Baker, P. and Egbert, J. (eds.). (2016). *Triangulating methodological approaches in corpus-linguistic research*. London: Routledge.
This volume brings together researchers who combine corpus linguistics with different methods from linguistics all to analyse the same data. It should prove to be an interesting read for those who want to explore how corpus methods might be combined with techniques of text analysis from their own discipline.

References

- Adolphs, S. and Carter, R. (2013). *Spoken corpus linguistics: From monomodal to multimodal*. London: Routledge.
- Aijmer, K. and Rühlemann, C. (2014). *Corpus pragmatics: A handbook*. Cambridge: Cambridge University Press.
- Aston, G. and Burnard, L. (1998). *The BNC handbook: Exploring the British national corpus with SARA*. Edinburgh: Edinburgh University Press.

- Baker, P. (2005). *Public discourses of gay men*. London: Routledge.
- Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.
- Baker, P. (2008). *Sexed texts: Language, gender and sexuality*. London: Equinox.
- Baker, P. (2009). The BE06 corpus of British English and recent language change. *International Journal of Corpus Linguistics* 14(3): 312–337.
- Baker, P. (2010). *Sociolinguistics and corpus linguistics*. Edinburgh: Edinburgh University Press.
- Baker, P., Brookes, G. and Evans, C. (2019). *The Language of Patient Feedback: A corpus linguistic study of online health communication*. London: Routledge.
- Baker, P. (2014). *Using corpora to analyze gender*. London: Bloomsbury.
- Baker, P. (2016). The shapes of collocation. *International Journal of Corpus Linguistics* 21(2): 139–164.
- Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T. and Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society* 19(3): 273–306.
- Baker, P., Gabrielatos, C. and McEnery, T. (2013). *Discourse analysis and media attitudes: The representation of Islam in the British press*. Cambridge: Cambridge University Press.
- Baker, P., Hardie, A., McEnery, T., Xiao, R., Bontcheva, K., Cunningham, H., Gaizauskas, R., Hamza, O., Maynard, D., Tablan, V., Ursu, C., Jayaram, B.D. and Leisher, M. (2004). Corpus linguistics and South Asian languages: Corpus creation and tool development. *Literary and Linguistic Computing* 19(4): 509–524.
- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing* 8(4): 243–257.
- Biber, D. (2011). Corpus linguistics and the study of literature: Back to the future? *Scientific Study of Literature* 1(1): 15–23.
- Biber, D. and Conrad, S. (2009). *Register, genre and style*. Cambridge: Cambridge University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. and Quirk, R. (1999). *Longman grammar of spoken and written English*. London: Longman.
- Brezina, V., McEnery, T. and Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20(2): 139–173.
- Brookes, G. and Baker, P. (2017). What does patient feedback reveal about the NHS? A mixed methods study of comments posted to the NHS Choices Online service. *BMJ Open* 7: e013821.
- Brookes, G. and McEnery, T. (2017). How to interpret large volumes of patient feedback: Methods from computer-assisted linguistics. *Social Research Practice* 4: 2–13.
- Carter, R. and McCarthy, M. (2017). Spoken grammar: Where are we and where are we going? *Applied Linguistics* 38(1): 1–20.
- Davies, M. (2009). The 385+ million word corpus of contemporary American English (1990 2008+): Design, architecture and linguistic insights. *International Journal of Corpus Linguistics* 14(2): 159–190.
- Davies, M. (2012). The 400 million word corpus of historical American English (1810 2009). In I. Hegedus and A. Fodor (eds.), *English historical linguistics 2010*. Amsterdam: John Benjamins, pp. 217–250.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1): 61–74.
- Durrant, P. and Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory* 6(2): 125–155.
- Ellis, N.C. and Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory* 5(1): 61–78.
- Ensslin, A. and Johnson, S. (2006). Language in the news: Investigations into representations of “Englishness” using *WordSmith Tools*. *Corpora* 1(2): 153–185.
- Fairclough, N. (1995). *Media discourse*. London: Edward Arnold.
- Firth, J.R. (1957). *Papers in linguistics 1934–1951*. Oxford: Oxford University Press.

- Flowerdew, L. (1997). Interpersonal strategies: Investigating interlanguage corpora. *RELC Journal* 28(1): 72–88.
- Flowerdew, L. (2012). *Corpora and language education*. Basingstoke: Palgrave Macmillan.
- Gablasova, D., Brezina, V. and McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning* 67(S1): 155–179.
- Gatto, M. (2014). *Web as corpus: Theory and practice*. London: Bloomsbury.
- Gregory, I., Donaldson, C., Murrieta-Flores, P. and Rayson, P. (2015). Geoparsing, GIS and textual analysis: Current developments in spatial humanities research. *International Journal of Humanities and Arts Computing* 9(1): 1–14.
- Hundt, M., Sand, A. and Siemund, R. (1998). *Manual of information to accompany the Freiburg-LOB corpus of British English ("FLOB")* [online]. Available at: www.hit.uib.no/icame/flob/index.htm.
- Hundt, M., Sand, A. and Skandera, P. (1999). *Manual of information to accompany the Freiburg-Brown corpus of American English ("Frown")* [online]. Available at: <http://khnt.hit.uib.no/icame/manuals/frown/INDEX.HTM>.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Jensen, K.E. (2014). Linguistics and the digital humanities: (computational) corpus linguistics. *Journal of Media and Communication Research* 30(57): 115–134.
- Johansson, S. (2007). *Seeing through multilingual corpora: On the use of corpora in contrastive studies*. Amsterdam: John Benjamins.
- Johansson, S., Leech, G. and Goodluck, H. (1978). *Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers*. Oslo: University of Oslo.
- Johnston, T. and Schembri, A. (2013). Corpus analysis of sign languages. In C. Chapelle (ed.), *Encyclopaedia of applied linguistics*. Oxford: Wiley-Blackwell, pp. 479–501.
- King, B.W. (2009). Building and analysing corpora of computer-mediated communication. In P. Baker (ed.), *Contemporary corpus linguistics*. London: Continuum, pp. 301–320.
- Knight, D. (2011). *Multimodality and active listenership: A corpus approach*. London: Bloomsbury.
- Knight, D., Adolphs, S. and Carter, R. (2014). CANELC – constructing an e-language corpus. *Corpora* 9(1): 29–56.
- Kučera, H. and Francis, W.N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.
- Laviosa, S. (2002). *Corpus-based translation studies: Theory, findings, applications*. Amsterdam: Rodopi.
- Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (ed.), *Directions in corpus linguistics: Proceedings of the nobel symposium 82, Stockholm, 4–8 August 1991*, pp. 105–122.
- Leech, G. (1997a). Introducing corpus annotation. In R. Garside, G. Leech and T. McEnery (eds.), *Corpus annotation: Linguistic information from computer text corpora*. London: Routledge, pp. 1–18.
- Leech, G. (1997b). Grammatical tagging. In R. Garside, G. Leech and T. McEnery (eds.), *Corpus annotation: Linguistic information from computer text corpora*. London: Routledge, pp. 19–33.
- Leech, G. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning* 50(4): 675–724.
- Leech, G., Hundt, M., Mair, C. and Smith, N. (2009). *Change in contemporary English: A grammatical study*. Cambridge: Cambridge University Press.
- Litchfield, I., Gale, N., Burrows, M. and Greenfield, S. (2016). Protocol for using mixed methods and process improvement methodologies to explore primary care receptionist work. *BMJ Open* 6: e013240.
- Love, R., Dembry, C., Hardie, A., Brezina, V. and McEnery, T. (2017). The spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22(3): 319–344.
- Mahlberg, M. (2013). *Corpus stylistics and Dickens's fiction*. London: Routledge.

- McEnery, T. (2005). *Swearing in English: Bad language, purity and power from 1586 to the present*. London: Routledge.
- McEnery, T. (2006). The moral panic about bad language in England, 1691–1745. *Journal of Historical Pragmatics* 7(1): 89–113.
- McEnery, T. and Baker, H. (2016). *Corpus linguistics and 17th-century prostitution: Computational linguistics and history*. London: Bloomsbury.
- McEnery, T. and Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
- McEnery, T. and Oakes, M. (1995). Sentence and word alignment in the Crater project: Methods and assessment. In S. Warwick-Armstrong (ed.), *Proceedings of the association of computational linguistics workshop SIG-DAT workshop*. Dublin: ACL, pp. 104–116.
- McEnery, T. and Wilson, A. (2001). *Corpus linguistics: An introduction*, 2nd ed. Edinburgh: Edinburgh University Press.
- Michel, J.-B., Shen, Y., Aiden, A., Veres, A., Gray, M., Google Books Team, Pickett, J., Holberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. and Aiden, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science* 331(6014): 176–182.
- Mills, S. (2003). *Gender and politeness*. Cambridge: Cambridge University Press.
- Nolte, M.I., Ancarno, C. and Jones, R. (2018). Inter-religious relations in Yorubaland, Nigeria: Corpus methods and anthropological survey data. *Corpora* 13(1): 27–64.
- Popper, K.R. [1935] (2006). *The logic of scientific discovery*. London: Routledge.
- Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics* 13(4): 519–549.
- Renouf, A. (2003). WebCorp: Providing a renewable data source for corpus linguists. In S. Granger and S. Petch-Tyson (eds.), *Extending the scope of corpus-based research: New applications, new challenges*. Amsterdam: Rodopi, pp. 39–58.
- Scott, M. (1997). PC analysis of key words – and key key words. *System* 25(2): 233–245.
- Scott, M. and Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sinclair, J. (1992). The automatic analysis of corpora. In J. Svartvik (ed.), *Directions in corpus linguistics: Proceedings of the nobel symposium 82, Stockholm, 4–8 August 1991*. Berlin: Mouton De Gruyter, pp. 379–397.
- Sinclair, J. (2003). *Reading concordances*. London: Longman.
- Smith, N., McEnery, T. and Ivanic, R. (1998). Issues in transcribing a corpus of children’s handwritten projects. *Literary and Linguistic Computing* 13(4): 312–329.
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.
- Widdowson, H. (2004). *Text, context, pretext: Critical issues in discourse analysis*. Oxford: Wiley-Blackwell.
- Wilson, A. and Thomas, J. (1997). Semantic annotation. In J. Svartvik (ed.), *Directions in corpus linguistics: Proceedings of the nobel symposium 82, Stockholm, 4–8 August 1991*. Berlin: Mouton De Gruyter, pp. 53–65.
- Xiao, R. and McEnery, T. (2004). A corpus-based two-level model of situation aspect. *Journal of Linguistics* 40(2): 325–363.
- Xiao, R. and McEnery, T. (2006). Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied Linguistics* 27(1): 103–129.

English language and classics

Alexandra Trachsel

Introduction

When dealing with the interaction of several research traditions, as, for instance, when exploring the mutual benefits gain from the combination of fields like digital humanities and other disciplines, it seems natural to reflect on one's own area of interest and to situate it in a larger context. As far as classics is concerned, if we follow the definition of the humanities recently given by Melissa Terras,¹ this subject obviously belongs to this domain of research.² But, among the many fields in the humanities, classics is one that focuses on the past, the way we can explore this heritage today and the influences these remote cultures, traditions and artefacts had and still have on subsequent human activities and thoughts. In this contribution, however, we would like to highlight a further aspect of classics which seems particularly relevant for reflections on the interactions between English, classics and digital humanities. This is the particularity of classics to be at the intersection of several languages. In this respect, the present contribution will focus on one facet of digital humanities, in so far that, in the projects that will be discussed, ancient texts and more specifically the languages, in which they were transmitted, are at the centre of interest. However, in the last section of our contribution we shall see how other research questions concerning classics, the focus of which lies beyond the texts themselves, can develop from projects that deal in the first place with ancient languages and their interaction with modern ones.

The choice to focus on this particular characteristic of classics and on the way in which people have reacted to it seems to be, in our opinion, a good starting point to illustrate one current research field where digital humanities and classics work together, moreover in relation with the English language. Furthermore, with this focus on projects that aim to improve our understanding of languages in a very broad sense, which include not only studies on their structures but also investigations into the ways in which two (or several) language systems may interact, and finally the development of the tools that would be needed to lower barriers between languages someone knows and those with which he or she is not familiar, we illustrate an aspect of classics that distinguishes such kind of studies from other fields of research related to antiquity, for instance, archaeology, and other domains that deal with artefacts other than texts or with human activities other than writing.³

Moreover, as none of the previously mentioned questions are restricted to the study of ancient languages, the issues such questions raise brings the field of classics closer to other disciplines of the humanities, especially those dealing with other languages, for instance, English. However, as classics deals mainly with historical languages, the scholars may be more concerned with providing translations of the ancient texts than in other fields that deal with still spoken languages. Or at least, it is nowadays very common that in research related to classics, scholars deal at least with two languages: the source language (the ancient language a scholar is interested in) and the target language (the modern language the scholar speaks).⁴ This is explained by the fact that there are now no native speakers of either ancient Greek or Latin who could use their intuition to speak about their own language or its literary productions.⁵ On the one hand, this situation means that there are scholars with different linguistic backgrounds who interpret the same ancient languages using their own grammars as a reference system. On the other, this situation also brings to the fore the question of the alignment of the languages involved, as often the ancient text is accompanied by a translation into a modern language to make its understanding easier for readers, who may not be well acquainted with the ancient language. As an answer to this situation, modern (bilingual) editions of ancient works tend, most of the time, to present the two texts involved (original and translation) as parallel texts in a side-by-side view, dividing each of them into sections that fit the space available on each page of the book. The division imposed is, therefore, external to the structure of either language, as it is defined by the format of the book. Consequently, as paragraphs have been defined to fit the page, each language keeps its inherent word order, which does not always match the structure of the second language. For readers who are not well-trained it is often difficult to find the exact wording of corresponding sentences, clauses, formulae or individual words, even if the meaning of the passage as a whole is understood without problems.⁶ A good example of a different way of proceeding is now provided by the new *Oxford Scholarly Editions Online*.⁷ This project takes advantage of digital tools to create a research environment that enhance its well-known and authoritative printed monolingual edition. A whole array of tools are utilised that also include, besides the critical apparatus from the printed version, an (English) translation of the text, lexical tools and further commentaries and notes from modern scholars. Here the notion of the page begins to disappear and the alignment between original text, its translation and any relevant comment on a given section of the text can be viewed for any part of the work in which a scholar may be interested. Nonetheless, the project keeps the indication of the pages as featured in the printed edition and also reproduces its layout in order that the navigation between the printed book and the online text is easier for scholars working in both environments. As far as the translation is concerned, it can be requested, if necessary, in a parallel column and it is then closely linked to the primary text, although not quite on the level of the individual words and their syntactic functions. Nevertheless, the user may easily navigate in both columns and have a side-by-side view of the section of the text in which he or she is interested in two languages (original and translation), along with a clear annotation system in both columns for the passages where further comments and aids are available.

However, the difficulty of aligning several languages is long-standing. It was also perceived in antiquity, as almost all ancient societies were at least bilingual, if not multilingual (Adams et al. 2002; Adams 2003). This situation triggered the need of combining several languages and stimulated all kinds of interrogations about the nature and structure of the languages involved. Such investigations started very early and developed in various regions of the Mediterranean Basin.⁸ But as far as Latin and ancient Greek is concerned, we may mention here Cicero (first century BCE) as one of the best-documented examples for these

two languages (Jocelyn 1973; Powell 1995). The most fitting example for our present discussion comes, however, from late antiquity. In the third century CE, Origen, one of the Christian Church fathers, composed the *Hexapla*, a work presenting six versions of the *Ancient Testament* aligned in six parallel columns (Markschies 2006; Fischer 2009: 133). The first two were in Hebrew (a Hebrew text and its transliteration into Greek letters) and the remaining four in Greek (the translations of Aquila and of Symmachus, followed by a corrected, annotated and revised version of the Septuagint and finally the translation of Theodotion [Fischer 2009: 133–138]). The purpose and the function of this work is still under discussion (Orlinsky 1936; Martin 2004) and only a few fragments have been preserved, the oldest of which being a fragment from the Cairo Genizah.⁹ Nonetheless, the work bears witness to the same issues that we face today when trying to find the best possible way to align two or more languages so that both can be understood in their grammatical, linguistic and semantic structures and studied together. Indeed, the means that Origen used to overcome the difficulty to match the texts from the two languages, which were neither written in the same writing system (Jewish scripts and Greek alphabet) nor in the same direction (Hebrew from right to left and Greek from left to right), was the splitting of the text into single words. Moreover, it was the Hebrew text which provided the grid for the structure of the whole book, as in the first column the Hebrew text was divided so that each line corresponded to only one single word (Martin 2004: 102). The subsequent columns were then adjusted to this word-by-word division as much as possible, so that the alignment was maintained in each column even if the readability of the subsequent columns suffered sometimes in this process (Norton 1998).

However, the Hebrew was not considered as the source language in the sense we would use it today. The text at the centre of Origen's scholarly activities was the Greek Septuagint (Schaper 1998). It is this text for which he wanted to provide a new edition freed from all changes, mistakes and additions introduced over the centuries.¹⁰ Origen thought that his goal would best be reached by providing several versions of the same text, some in another language, despite the relatively simple means that he had at his disposal to enact the alignment of the linguistic sequences.

Today the means to compare different languages in their syntactical behaviours are far more sophisticated and, in particular, the so-called dependency grammar and its digital development known as treebanking are powerful tools to carry out such analyses. A first example – which perfectly illustrates the interaction between old research questions and new methods, tools and know-how coming from digital humanities – is the University of Oslo's PROIEL project.¹¹ Moreover, this project also comes very close to Origen's example from antiquity, as several versions of the *New Testament*, each in a different language,¹² are studied in parallel using treebanking annotations. The scope of the modern project is quite different from the ancient one, as it is driven by linguistic research questions with a focus on the grammatical and syntactic behaviours of the different languages involved. In antiquity, Origen's project, while comparing several versions of one text, aimed to provide the best and most accurate rendering of this same text. It was therefore a more philological and theological project. Nonetheless, comparing both undertakings is justified, as some issues are common to both. For instance, the previously mentioned difficulty of dividing the texts into comparable sequences is again addressed by the Oslo project. Here the scholars decided to use sentence division as a first guiding principle before proceeding, in a second step, to a finer grained division, which segmented each sentence unity into its elements (e.g. words, numbers or punctuation marks). However, they noted from the beginning that this procedure created difficulties. Some of the ancient versions of the *New Testament* (the Latin Vulgata

for instance) had no proper and consistent sentence divisions and compromise has had to be found to provide this information (Haug et al. 2009: 24–25). Nonetheless, analysing the languages in accordance with the format of dependency grammar allowed the scholars to highlight their similarities, as they all were Indo-European languages, despite their apparent differences and distance in time and space. On the other, however, it also allowed the researchers to show where the differences between the six languages involved lay and how the tool used for the analyses had to be improved to show all the particularities of each language. Finally, the scholars also used the results of their annotations to better describe the common linguistic features that distinguish the group of Indo-European languages, to which the six languages studied belong, from other non-Indo-European languages.

However, as far as English is concerned, this language seems not to have had a prominent position in the PROIEL project. It did not feature among the languages analysed (namely ancient Greek, Latin, Armenian, Gothic and Church Slavonic), which were purposely chosen for the corpus of texts that the project wanted to study.¹³ Likewise, little information can be gained about the role English may have played in other phases of the project, for instance with regard to the language proficiency of the contributors or how this aspect may have shaped their research questions. The contributors were chosen in accordance with their proficiency in the languages studied (Haug et al. 2009: 33), and we learn little about the influence their different linguistic backgrounds may have had on their analyses of the sentences or on how the usage of an English terminology for the studied phenomena may have influenced their perception of the different languages.¹⁴ To obtain further information regarding this aspect, we may mention briefly some projects focusing on modern languages, before returning to other treebanking projects in the field of classics.

Background

Our first example here should certainly be the pioneering Prague Czech Dependency Treebank, as this project not only involves the English language but also shows how the experience and developments from the field of digital humanities – in particular, from computational linguistics – enhance the scope of a study and pushed forward the understanding of research questions that are linked to it.¹⁵ This project combined the digitised and annotated text corpus from the *Penn Treebank Project* with the experience from the Prague Dependency Grammar (Čmejrek et al. 2004) and closely described the transformations, which were necessary to adapt the two annotation systems and the structural differences that scholars encountered during this process (Cinková et al. 2008). As far as the status of the two languages, and in particular the one of English, is concerned, the project began from the annotated corpus of English texts provided by the Penn Treebank (taken from the *Wall Street Journal*) and was translated into Czech before annotating the translation. This procedure seems at first sight to prioritise the English language, but the project had also to adapt the Penn Treebank annotation style and its phrase trees to the Prague Dependency system. Therefore, the Czech tradition of annotation, and the Czech language on which it was first used, was also well represented and influenced the way in which the English language was understood and described. We have, therefore, in this project a well-functioning interaction of two languages and their linguistic traditions, so that a good balance was established, and this not only between the two languages but also between a more traditional expertise and newer approaches gained from the opportunity offered by digital methodologies and tools.

This project became paradigmatic and, in the subsequent years, it inspired other projects, for instance the *ParDeepBank*, a collaborative project of five research institutions,¹⁶

intended to provide a parallel treebank for Bulgarian, Portuguese and English. For this project, too, the Penn Treebank corpus was the starting point. The texts from the Penn Treebank were translated into the other languages and then annotated so that the syntactical structure of each language could be matched with the others. With regard to our approach here, this project is of particular interest as the Bulgarian language has another script, as, for instance, the Hebrew when compared with ancient Greek in Origen's project. Therefore, similar questions of transliteration from words written in one language (English) to the script of the other (Bulgarian) were involved, but they can now be handled in a new research environment, which not only provides the means to better understand the processes involved. It also offers the possibility to express the results more accurately and in a more sustainable way, so that they remain easily available for subsequent research.

Indeed, in close connection with these projects, other initiatives which built on the experience from the *Penn Treebank Project* were developed, for instance the so-called *Universal Dependencies* (UD) (Nivre et al. 2016). This international annotation system emerged from the combination of several other projects. Over the years, it absorbed all previous experiences into one single effort, the scope of which exceeded those of each individual project. First, the *Stanford Dependencies* (SD) (de Marneffe et al. 2006, 2014) was merged with, and expanded by, several other tagsets and annotations schemes, among which the one developed by the previously mentioned Prague Dependency Grammar, the so-called *HamleDT* normalisation tool (Zeman et al. 2012). At this point, promising experiences with the alignment of six languages (English, Swedish, German, French, Spanish and Korean) were made (McDonald et al. 2013). The project was then again expanded under the *Universal Stanford Dependencies* (USD), which, under its latest title (*Universal Dependencies*), integrates now more than 40 languages. The choice of the six original languages is again revealing: from very early this project included languages, which came from beyond the Indo-European group (e.g. Korean). In so doing, it expands for instance the scope of the PROIEL project mentioned earlier, which focused on Indo-European languages and described the characteristic behaviours of this group, in order to better understand the differences with regard to non-Indo-European languages. Nevertheless, the scholars were well aware that, despite their global approach, the English language was still the starting point, as the first SD was made for the English language and the later developments, including other languages, attempted to keep their annotations system as close as possible to the original English SD representation (de Marneffe et al. 2014: 1). However, the later developments, through the expansion to the *Universal Stanford Dependencies* (USD) first and then to the current *Universal Dependencies* (UD), show the willingness of the scholars involved to overcome this heritage and to keep the balance between the different languages and their heterogeneous perception of morphological, grammatical and syntactical phenomena. The most recent developments of this project and its even more global approach suggest that the analysing tool may have the potential to reach beyond the individual languages and to encompass almost all known languages. With this in mind, we may now turn to the project in classics, and especially to the *Ancient Greek and Latin Dependency Treebank* (AGLDT), and investigate how these issues are dealt with in the field of classics and how the mentioned project contributes to the more global approach we have just described.

Current contributions and research

In the field of classics, treebanking projects began in 2006, first for Latin and then also for ancient Greek. In its first phase, the AGLDT was developed in connection with the *Perseus*

Digital Library (Tufts University) and its infrastructure, with the goal to provide an annotation system for the two ancient languages (Bamman and Crane 2010: 543). It was initially based on the Prague Dependency annotation system (Bamman 2014: 436), but is now part of the UD.¹⁷ However, since the beginning of this initiative, the differences between the historical treebanks of these two ancient languages and those established for modern languages have been highlighted (Bamman et al. 2009; Bamman and Crane 2011). First, as mentioned in the introduction, there are no native speakers of these languages who could rely on their intuition when analysing the sentences. Therefore, the speed of the annotations was, for instance, much slower than when done by native speakers. But the slowness of the annotators was not only due to their levels of language proficiency. Their hesitations with regard to giving one correct analysis of a sentence can also be explained by the fact that the language registers in the analysed corpora were quite different. The preserved texts from antiquity were highly stylised and rhetorically and/or artistically elaborated and triggered a long tradition of annotations, interpretations and commentaries, whereas for the Penn Treebank, for instance, extracts from the *Wall Street Journal* were taken. The language register in this English corpus is therefore much more representative of the language spoken by modern annotators than would be any of the ancient texts that were preserved. Because of the highly stylised language register of the two ancient languages, and also because of the remoteness of the topics dealt with in these texts, intense and continuous study was undertaken to understand this heritage from antiquity and to bring it closer to a modern audience. Therefore, scholars who developed the treebank system for ancient languages had to adapt the available procedure of annotation in order to document the semantic richness of the two languages, and two peculiarities characterise now the historical treebanks of ancient Greek and Latin.

The first is the development of the so-called ‘scholarly treebanks’ (Bamman et al. 2009: 5–9), which functions as a complement to the ‘standard treebank’ known for modern languages. These ‘scholarly treebanks’ were an answer to the additional difficulties, which annotators experienced when analysing ancient texts. Indeed, not only did the highly stylised register of the language in these texts and their remoteness in time from the linguistic background of the annotators create difficulties, but additional problems arose from the fact that, because of the long tradition of manual copying each work to preserve it, the ancient texts were often transmitted in several versions. In many cases, we find, therefore, textual variants in the wording of some of the transmitted sentences, or for some of their parts (such as clauses, words or even letters), which may influence either the syntactic analysis or even the meaning of these sentences. In such cases, it is highly relevant to maintain the ambiguities and to document the divergences and disagreements among annotators. Thus, in opposition to the standard form of treebanking, where a harmonisation among the annotators is sought, these ‘scholarly treebanks’ keep the traces of the individual interpretations of the scholar who provided the annotations. This procedure offers the opportunity to take account of several previous scholarly decisions regarding a text, which may have influenced an annotator, and this may concern either the wording of the text (editorial decisions) or the meaning of the content (literary interpretations).

Second, to give the annotators access to this rich tradition of studies concerning the corpus of ancient texts, the AGLDT was embedded in the research environment of the *Perseus Digital Library*, which provided a great variety of contextual material such as secondary publications, morphological annotations, links to lexica and dictionaries and information about variant manuscript readings (Bamman and Crane 2010: 544). Both of these characteristics of the historical treebanks developed in the field of classics have some relevance, not only with regard to the role the English language may have played in the project but also

with regard to how the experience gained from the annotation of this specific corpus of texts can serve the development of further tools and research questions.

Critical issues and topics

First, we can focus on the environment offered by the *Perseus Digital Library* (Bamman and Crane 2011: 84–90). This project, hosted since its beginning in the mid- 1980s, at Tufts University, emerged from an English-speaking environment. Due to this context, the English language is clearly associated with the project and stands out among the languages represented, even though the digital library hosts several collections of texts in other languages (e.g. Arabic, old Norse and Renaissance Italian, beside Latin and ancient Greek).¹⁸ This situation is easily explained: if modern translations are provided, they are English translations. The same applies to the lexica and dictionaries that are available among the interpretation tools of the ancient languages. They provide a translation of the Greek or Latin lemma found in the original text into English equivalent terms.¹⁹ The collaborators of the project are well aware of this bias and have tried since early on to enlarge the focus by associating with the project as many contributors from other linguistic backgrounds as possible in order to allow a greater diversity in the approach and to take account of the different national traditions in the study of the ancient languages. Second, the development of such projects as the AGLDT has also increased the global perspective of the *Perseus Digital Library*. This will be revisited more specifically later. Finally, the latest evolution of this digital library especially demonstrates the willingness of researchers implementing this project to embrace a truly global approach (Berti et al. 2014). For the moment, however, these newest aspects of the project, developed as collaboration between the *Perseus Digital Library* and the chair of digital humanities at the University of Leipzig, are focusing on ancient Greek and Latin and Persian literature.²⁰ The scope of the two parts of this initiative are, however, very different: the aim of the one dealing with ancient Greek and Latin consists in assembling, in a comprehensive open-source library, all preserved source texts of ancient Greek and Latin. This means that the focus is on the preserved physical items that bear the different texts, which will be kept in their original languages (e.g. the medieval manuscripts as well as inscriptions, papyri, ostraca and other written artefacts, Berti et al. 2014). Future development will therefore also go beyond the present focus on texts as objects of research and add information regarding other objects from antiquity or links to further databases with relevant information about the provenance of the artefacts, their state of preservation and their current location, or the places where they were found.

On the contrary, in the *Open Persian Project*, the aim is to give to an audience as large as possible access to Persian literature. Here the translations of the texts in several languages are at the centre, and the alignment of languages will play a fundamental role. This brings us back to our topic of treebanking and to its focus on texts as objects of research. The project is closely related to the AGLDT, from where we started, and benefits from its experience, but exports it into another field of research. This allows us now to leave the *Perseus Digital Library* and to turn more specifically to the AGLDT and to analyse what status English has within this project.

First, like the previously mentioned SD, the project began in the United States. In a group of six American Universities, students were trained in the ‘standard treebanking’.²¹ Therefore, the English language again seems at first sight to be among the driving factors of the digital turn to which the AGLDT contributed. This observation is also true for the latest development of the project, which aims to add a further layer of annotations, the so-called

semantic annotation (Gelano 2015). For this aspect, too, the project takes an English starting point, as the English standard Grammar for ancient Greek (Smyth 1920) is used as reference work for this additional semantic layer of annotations. Yet again scholars worldwide were associated with each of the different developments of the project, so that the linguistic backgrounds of the annotators were highly diversified (Bamman et al. 2009: 545; Bamman and Crane 2011: 4; Gelano 2015: 30). Moreover, the AGLDT is now part of the UD and contributes to its double aims, to create a universal annotations system, which would be able to define the specificities of each language (Nivre 2016: 1663).

How this ambitious goal might be reached can be illustrated with two examples of projects from classics that build on the experience of the AGLDT. They will also show how the approach developed for this annotation project open up other research questions.

Future directions

The first project to be mentioned here, the *Hellespont Project* hosted at the *Deutsches Archäologisches Institute* (DAI),²² will reveal how more specific research can be undertaken, even if the cross-linguistic characteristics of the annotation scheme of the treebanks are maintained. This initiative builds upon the annotated corpus provided by the AGLDT and aims to add a further layer of annotation that goes beyond the linguistic level and highlight the events and the actors mentioned in a text (semantic-pragmatic information). This information is highly relevant for research in the field of classics that deal with other objects than the transmitted texts, their forms and structures. In the case here of the *Hellespont Project*, the additional layer of information provides, for instance, data for the analysis of historical events in a given time span and for the appreciation of ancient actors and their deeds, either from an ancient point of view (e.g. their emblematic function) or from the modern perspective of reception studies.

However, when starting from the annotated treebank corpus rather than from any other edition of the ancient texts for such kinds of research, the scholars may benefit from a significant advantage. The information gathered from a corpus that is annotated with such a semantic-pragmatic layer can also be expressed in other languages than the original one, for instance, in English, for a larger audience of experts than up to now. Additionally, the results of such studies can also be displayed and visualised in a different form, so that other aspects may be highlighted that would not have been possible to see in a more traditional form of release. Moreover, such projects also require the contributions of other fields of digital humanities than those focusing on the computational analysis of languages. For instance, when visualisation is involved, the focus is as much on the transformation of the information through the process of visualisation as on the release of the results.

This is similar with a second project. Currently developed at the University of Nebraska-Lincoln, this initiative uses the experience from the AGLDT to deal with textual reuses (Gorman and Gorman 2016). This is a highly relevant issue for studies in classics and touches on the question about the access we have to objects of research from antiquity. As a great number of texts were not preserved and we have only indirect access to extracts from them, through quotations made by other ancient readers, the way ancient scholars reused works from predecessors is of crucial importance. The project under discussion here focuses on this research question and tries to use the syntactical information provided by the treebanks to find recurrent syntactic patterns in a corpus of texts. These patterns could then be used as criteria to identify the writing style of a given author, for instance, when the authorship of a passage is not securely established. This is often the case, when modern scholars have to

make hypotheses about which segment in a preserved text could be a quotation or textual reuse from a lost author. With this research question, the project touches a larger domain of investigation, which is explored by other projects involved in digital humanities and joins the common effort to tackle the issues from a new angle. Two of these, the *e-AQUA Project* and the *Perseid*, are being developed at the University of Leipzig.²³ Both use digital methods and tools to establish greater accuracy in our treatment of ancient textual reuses. Moreover, both work in close collaboration with projects focusing on textual reuses in modern languages, bridging in doing so the gap between classics with its remote past and our present time.

In addition to these interdisciplinary approaches, the Nebraska project, by working on data obtained by using the information from treebanking annotations, provides the opportunity to share more easily the results achieved and to combine and exchange the expertise gathered in the individual projects. It would, for instance, be possible to zoom in, when the focus of the research is on a given language or on a set of languages and to define the specificities of textual reuses in each language involved, so that issues related to each field of research, for instance, those linked to the question of the transmission of lost works in classics, can be addressed. However, with this approach it would also be possible to highlight the common features of a larger group of languages, so that the phenomenon can be studied from a more global perspective and with regards to the interaction of different cultures.

The latest trends, which these projects seem to indicate, may then provide the link back the beginning of our contribution and to antiquity. The goal these projects have and the methods and tools they develop show us how old questions, already raised in antiquity and traceable in works like Origen's *Hexapla* (manual alignment of a text available in different languages) can now be investigated with techniques that pave the way for a truly global approach (UD). They provide the methodology and the means to deepen our understanding of human languages and to search for answers to present-day issues, such as the understanding of different cultures and the lowering of linguistic barriers.

Notes

- 1 Terras 2016: 1736: we defined the humanities as 'academic disciplines that seek to understand and interpret the human experience, from individuals to entire cultures, engaging in the discovery, preservation, and communication of the past and present record to enable a deeper understanding of contemporary society'.
- 2 See Rosenbloom 2012, for the special shape that the inclusion of classics among the disciplines of the humanities gives to the interaction between information technologies in a broader sense and this specific domain of research.
- 3 This is another characteristic of classics, although it is more relevant when comparing this field of research with those focusing on other (modern) languages. As the transmitted corpus of texts contains very little evidence about the spoken form of the languages, classics is, most of the time, confined to work with their written forms.
- 4 We may exclude for the time being the long tradition of Latin as a scholarly language used by scholars to speak about ancient texts. Despite the fluency the scholars achieved, it remained a non-native language, which was learned, to different levels of proficiency, in schools and was expected to revive a highly artificial literary level of Latin.
- 5 Here again we have to qualify our statement. Even if there is an amazing continuity in the Greek language between the form preserved by the earliest witnesses and the one currently spoken, there are many factors that modified the language over time, so that modern Greek is not immediately considered as the same language as ancient Greek. This distinction may be illustrated by the division established by *The Thesaurus Linguae Graecae* at University of California, Irvine (<http://stephanus.tlg.uci.edu/>). Their corpus comprises texts written since the earliest times until the fall

of Constantinople (in 1453 CE). In our study, modern Greek belongs to the same category as other modern languages, which are currently used to speak about ancient texts.

- 6 The Loeb classical library is a good example here (see at www.loebclassics.com/), although the situation is similar with other bilingual editions when they are released in form of printed books.
- 7 www.oxfordscholarlyeditions.com/classics.
- 8 For some examples, see Adams et al. 2002 and for a larger Mediterranean context, Galter 1995: 25–50; Porter 1995: 51–72; Seminara 2012.
- 9 T-S 12.182 (dated to the seventh century CE) also designed as Rahlfs 2005. An image of the document can be found at <http://cudl.lib.cam.ac.uk/view/MS-TS-00012-00182/1>. For more details about the decipherment, see Jenkins 1998. The most important witness is, however, the so-called Mercati Fragment (alternatively referred to as Rahlfs 1098) probably dated to the tenth century. For a description of this fragment, see Flint 1998.
- 10 According to legend, the Septuagint was made in the third century BCE, but over the centuries until Origen's time in the third century CE, much was changed and there was no standard version. See Kreuzer 2011 and Wandrey 2006 for a short overview.
- 11 www.hf.uio.no/ifikk/english/research/projects/proiel/ and Haug et al. 2009 for a detailed description of the project.
- 12 Ancient Greek, Latin, Armenian, Gothic and Church Slavonic (see Haug and Jöhndal 2008: 27).
- 13 The aim of the project was a comparative study of the oldest extant versions of the *New Testament* in Indo-European languages.
- 14 For the history of the terminology, see Marcus et al. 1993 and Čmejrek et al. 2004.
- 15 See Hajičová 2011 for a summary on the larger research context, at the intersection between linguistics and computational linguistics, in which the project has been developed. For the relation between digital humanities, computational linguistics and other forms of computing in the humanities, see Svensson 2009.
- 16 The project was a collaboration of Stanford University, the DFKI GmbH, the Humboldt University of Berlin, the University of Lisbon and the Bulgarian Academy of Science. See further Flickinger et al. 2012.
- 17 www.dh.uni-leipzig.de/wo/projects/ancient-greek-and-latin-dependency-treebank-2-0/ and <http://universaldependencies.org/grc/overview/introduction.html>.
- 18 For the word counts of the Greek and Latin collection, see www.perseus.tufts.edu/hopper/collection?collection=Perseus:collection:Greco-Roman. This over-representation of English, however, is also visible in the *Oxford Scholarly Editions Online*. Here, too, the English language is prioritised by the choice the editors make among the printed books that are made available through the online version. For a full list of the works included in the project, see www.oxfordscholarlyeditions.com/page/18/title-lists.
- 19 For the ancient Greek this is most prominently the LSJ (H.G. Liddell, R. Scott and H.S. Jones, *A Greek-English Lexicon*, Oxford 1996) and for the Latin, the Lewis & Short (C.T. Lewis and C. Short, *A New Latin Dictionary*, New York/Oxford 1879).
- 20 www.dh.uni-leipzig.de/wo/open-philology-project/.
- 21 Tufts University, Brandeis University, the College of the Holy Cross, Furman University, the University of Missouri at Kansas City and the University of Nebraska at Lincoln.
- 22 <http://hellespont.dainst.org> and Mambrini 2016.
- 23 www.eaqua.net/ and www.dh.uni-leipzig.de/wo/perseids-fragmentary-texts-editor/.

Further reading

- 1 Celano, G.G.A. and Crane, G. (eds.). (2016). *Topical issue on treebanking and ancient languages: Current and prospective research* [special issue]. *Open Linguistics* 2.1.
The most recent volume on treebanking projects involving ancient Greek and Latin. Together with the previously mentioned contribution (Gorman and Gorman 2016), other promising usages of the treebanking annotation are presented.
- 2 Dickinson, M. et al. (eds.). (2015). *Proceedings of the fourteenth international workshop on treebanks and linguistic theories (TLT14)*. Warsaw.
This volume contains contributions from several treebanking projects, some based in classics, some dealing with a great variety of modern languages.

- 3 Almas, B. and Berti, M. (2013). Perseids collaborative platform for annotating text re-uses of fragmentary authors. In F. Tomasi and F. Vitali (eds.), *DH-Case 2013. Proceedings of the 1st international workshop on collaborative annotations in a shared environment: Metadata, vocabularies and techniques in the digital humanities*. ACM New York: ACM.

This contribution provides some additional information about the Perseids-Fragmentary Texts Editor.

References

- Adams, J.N. (2003). *Bilingualism and the Latin language*. Cambridge: Cambridge University Press.
- Adams, J.N., Janse, M. and Swain, S. (eds.). (2002). *Bilingualism in ancient society: Language contact and the written text*. Oxford: Oxford University Press.
- Bamman, D. (2014). Dependency grammar and Greek. In G.K. Giannakis (ed.), *Encyclopedia of ancient Greek language and linguistic*. Leiden: Brill, pp. 434–438.
- Bamman, D. and Crane, G. (2010). Corpus linguistics, treebanks and the reinvention of philology. In K.-P. Fähnrich and B. Franczyk, B. (eds.), *Informatik 2010. Service Science – Neue Perspektiven für die Informatik*. Bonn: Gesellschaft für Informatik e.V., pp. 541–551.
- Bamman, D. and Crane, G. (2011). The ancient Greek and Latin dependency treebank. In C. Sporleder, A. van den Bosch and K. Zervanou (eds.), *Language technology for cultural heritage: Selected papers from the LaTeCH workshop series*. Berlin and Heidelberg: Springer-Verlag, pp. 79–98.
- Bamman, D., Mambri, F. and Crane, G. (2009). An ownership model of annotation: The ancient Greek dependency treebank. In *TLT 2009: Proceedings of the eighth international workshop on treebanks and linguistic theories conference*. Milan Italy: Northern European Association for Language Technology (NEALT), 2009–11. Tartu.
- Berti, M., Franzini, G., Franzini, E., Celano, G.G.A. and Crane, G.R. (2014). L'Open Philology Project dell'Università di Lipsia. Per una filologia 'sostenibile' in un mondo globale. In M. Agosti and F. Tomasi (eds.), *Collaborative research practices and shared infrastructures for humanities computing*. Padova: Coop. Libreria Editrice Università di Padova, pp. 151–162.
- Cinková, S., Hajičová, E., Panevová, J. and Sgall, P. (2008). The tectogramatics of English: On some problematic issues from the viewpoint of the Prague dependency treebank. In J. Nivre, M. Dahlöf and B. Megyesi (eds.), *Resourceful language technology: Festschrift in honor of Anna Săgvall Hein*. Uppsala: Acta Universitatis Upsaliensis, pp. 33–48.
- Čmejrek, M., Cuřín, J. and Havelka, J. (2004). Prague Czech-English dependency treebank: Any hopes for a common annotation scheme? In A. Meyers (ed.), *HLT-NAACL 2004 workshop: Frontiers in corpus annotation*. Boston, pp. 47–54.
- de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J. and Manning, C.D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In N. Calzolari et al. (eds.), *Proceedings of 9th international conference on language resources and evaluation (LREC 2014)*. Reykjavik, pp. 4585–4592.
- de Marneffe, M.-C., MacCartney, B. and Manning, C.D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC-06*. Genoa, pp. 449–454.
- Fischer, A.A. (2009). *Der Text des Alten Testaments. Neubearbeitung der Einführung in die Biblia Hebraica von Ernst Würthwein*. Stuttgart: Deutsche Bibelgesellschaft.
- Flickinger, D., Kordoni, V., Zhang, Y., Branco, A., Simov, K., Osenova, P., Carvalheiro, C., Costa, F. and Castro, S. (2012). ParDeepBank: Multiple parallel deep treebanking. In I. Hendrickx, S. Kübler and K. Simov (eds.), *Proceedings of the 11th international workshop on treebanks and linguistic theories*. Lisbon, pp. 97–108.
- Flint, P.W. (1998). Columns I and II of the Hexapla: The evidence of the Milan Palimpsest. In A. Salvesen (ed.), *Origen's Hexapla and fragments*. Tübingen: Mohr Siebeck, pp. 125–132.
- Galter, H.D. (1995). Cuneiform bilingual royal inscriptions. In S. Izre'el and R. Drory (eds.), *Israel oriental studies 15: Language and culture in the near East*. Leiden: Brill, pp. 25–50.

- Gelano, G.G.A. (2015). Semantic role annotation in the ancient Greek dependency treebank. In M. Dickinson et al. (eds.), *Proceedings of the fourteenth international workshop on treebanks and linguistic theories (TLT14)*. Warsaw, pp. 26–34.
- Gorman, V.B. and Gorman, R.J. (2016). Approaching questions of text reuse in ancient Greek using computational syntactic stylometry. In G.G.A. Celano and G. Crane (eds.), *Topical issue on treebanking and ancient languages: Current and prospective research* [special issue]. *Open Linguistics* 2.1: 500–510.
- Hajičová, E. (2011). Computational linguistics without linguistics? View from Prague. *Linguistic Issues in Language Technology* 6(6): 1–20.
- Haug, D.T.T. and Jøhndal, M.L. (2008). Creating a parallel treebank of the old Indo-European Bible translations. In C. Sporleder and K. Ribarov (eds.), *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, pp. 27–34.
- Haug, D.T.T., Jøhndal, M.L., Eckhoff, H.M., Hertzberg, M.J. and Muth, A. (2009). Computational and linguistic issues in designing a syntactically annotated parallel corpus of Indo-European languages. *Traitement automatique des langues* 50(2): 17–45.
- Jenkins, R.G. (1998). The first column of the Hexapla: The evidence of the Milan Codex (Rahlf's 1098) and the Cairo Geniza fragment (Rahlf's 2005). In A. Salvesen (ed.), *Origen's Hexapla and fragments*. Tübingen: Mohr Siebeck, pp. 88–102.
- Jocelyn, H.D. (1973). Greek poetry in Cicero's prose writing. *YCS* 23: 61–111.
- Kreuzer, S. (2011). Entstehung und Entwicklung der Septuaginta im Kontext alexandrinischer und frühjüdischer Kultur und Bildung. In M. Karrer and W. Kraus (eds.), *Septuaginta Deutsch. Erläuterungen und Kommentare zum griechischen Alten Testament*. Stuttgart: Deutsche Bibelgesellschaft, pp. 3–39.
- Lewis, C.T. and Short, C. (1879). *A new Latin dictionary*. New York and Oxford: Clarendon.
- Liddell, H.G., Scott, R. and Jones, H.S. (1996). *A Greek-English Lexicon*, 9th ed. Oxford: Clarendon.
- Mambrini, F. (2016). Treebanking in the world of Thucydides: Linguistic annotation for the Hellenist project. *Digital Humanities Quarterly* 10(2): 1–10.
- Marcus, M.P., Santorini, B. and Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics* 19(2): 313–330.
- Markschies, C. (2006). Origenes [2]. In H. Cancik and H. Schneider (eds.), *Brill's New Pauly*. Brill Online. First published online. Consulted online on 6 January 2017. Available at: http://dx.doi.org.brillspnewpauly.emedien3.sub.uni-hamburg.de/10.1163/1574-9347_bnp_e900660.
- Martin, M.J. (2004). Origen's theory of language and the first two columns of the Hexapla. *HTR* 97: 99–106.
- McDonald, R., Nivre, J., Quirnbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N. and Lee, J. (2013). Universal dependency annotation for multilingual parsing. In P. Fung and M. Poesio (eds.), *Proceedings of the 51st annual meeting of the association for computational linguistics*. Sofia, pp. 92–97.
- Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R. and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of 10th international conference on language resources and evaluation (LREC 2016)*. Portorož, pp. 1659–1666.
- Norton, G.J. (1998). Observations on the first two columns of the Hexapla. In A. Salvesen (ed.), *Origen's Hexapla and fragments*. Tübingen: Mohr Siebeck, pp. 103–124.
- Orlinsky, H.M. (1936). The columnar order of the Hexapla. *JQR* 27: 137–149.
- Porter, B.N. (1995). Languages, audience and impact in imperial Assyria. In S. Izre'el and R. Drory (eds.), *Israel oriental studies 15: Language and culture in the near East*. Leiden: Brill, pp. 51–72.
- Powell, J.G.F. (1995). Cicero's translations from Greek. In J.G.F. Powell (ed.), *Cicero the philosopher: twelve papers*. Oxford: Clarendon, pp. 273–300.
- Rosenbloom, P.S. (2012). Towards a conceptual framework for the digital humanities. *Digital Humanities Quarterly* 6(2): 1–8.
- Schaper, J. (1998). The origin and purpose of the fifth column of the Hexapla. In A. Salvesen (ed.), *Origen's Hexapla and fragments*. Tübingen: Mohr Siebeck, pp. 3–15.

- Seminara, S. (2012). The most ancient translations: The case of Mesopotamia. In *The study of Asia – between antiquity and modernity*. Cagliari: Coffee Break Project, pp. 71–74.
- Smyth, H.W. (1920). *A Greek grammar for colleges*. New York: American Book Co.
- Svensson, P. (2009). Humanities computing as digital humanities. *Digital Humanities Quarterly* 3(3): 1–17.
- Terras, M. (2016). A decade in digital humanities. *Journal of Siberian Federal University: Humanities & Social Sciences* 7(9): 1637–1650.
- Wandrey, I. (2006). Septuagint. In H. Cancik and H. Schneider (eds.), *Brill's New Pauly*. Brill Online. First published online. Consulted online on 6 January 2017. Available at: http://dx.doi.org/brills.newpauly.emedien3.sub.uni-hamburg.de/10.1163/1574-9347_bnp_e1109300.
- Zeman, D., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z. and Hajič, J. (2012). HamleDT: To parse or not to parse?. In N. Calzolari et al. (eds.), *Proceedings of the eight international conference on language resources and evaluation (LREC'12)*. Istanbul, pp. 2735–2741.

Internet references

- Ancient Greek and Latin Dependency Treebank, University of Leipzig and Tufts University. Available at: www.dh.uni-leipzig.de/wo/projects/ancient-greek-and-latin-dependency-treebank-2-0/.
- Universal Dependencies v2, Ancient Greek. Available at: <http://universaldependencies.org/grc/overview/introduction.html>.
- e-AQUA, University of Leipzig. Available at: www.eaqua.net/.
- Open Philology Projects, University of Leipzig. Available at: www.dh.uni-leipzig.de/wo/open-philology-project/.
- Oxford Scholarly Editions Online. Available at: www.oxfordscholarlyeditions.com/classics.
- PROIEL-Project, University of Oslo. Available at: www.hf.uio.no/ifikk/english/research/projects/proiel/.
- Perseids-Fragmentary Texts Editor, University of Leipzig. Available at: www.dh.uni-leipzig.de/wo/perseids-fragmentary-texts-editor/.
- The Thesaurus Linguae Graecae, University of California, Irvine. Available at: <http://stephanus.tlg.uci.edu/>.
- The Digital Loeb Classical Library. Available at: www.loebclassics.com/.
- The Cambridge Digital Library. Available at: <http://cudl.lib.cam.ac.uk/view/MS-TS-00012-00182/1>.
- The Hespont Project, Deutsches Archäologisches Institut. Available at: <http://hellespont.dainst.org>.
- The Perseus Digital Library, Tufts University. Available at: www.perseus.tufts.edu/hopper/collection?collection=Perseus:collection:Greco-Roman.

English language and history

Geographical representations of poverty in historical newspapers

Ian N. Gregory and Laura L. Paterson

Introduction

The use of computational approaches in history is not new (Boonstra et al. 2004). However, until fairly recently, their use has been restricted to a relatively small number of fields such as economic history and historical demography. These are fields that make extensive use of quantitative sources, typically in tabular form, that have long been well-suited to analysis using a computer. Despite the potential benefits of quantitative sources, the overwhelming majority of historians use textual sources, and thus, computational approaches have traditionally had little penetration into mainstream history. Over the past decade or so this has changed rapidly. The mass digitisation and dissemination of historical source material, such as Early English Books Online (EEBO), Eighteenth Century Collections Online (ECCO), the British Library Newspaper Collections, *The Times* Digital Archive, Project Gutenberg and many others, have meant that large amounts of historical texts are now available in digital form (Hitchcock 2013). Thus, historians now have access to an unprecedented volume of digitised source material. However, the computational techniques used to analyse such materials remain rooted in traditional close reading, with the digital simply providing ease of access and the use of keyword searching. There is thus a requirement to adopt and adapt the digital approaches available from other fields, and develop new techniques where appropriate, such that the scale of the resources available to us can be exploited in a way that is impossible for close reading alone. This challenge is not unique to history; it has occurred across humanities disciplines, providing a major impetus to the field of digital humanities (Kirschenbaum 2010; Schreibman et al. 2014; Terras 2016), where a major theme is the use of digital approaches to the analysis of texts in disciplines, including English literature (Moretti 2013; Jockers 2013) and History (Atkinson and Gregory 2017; Pumfrey et al. 2012).

One of the key, and under-represented, opportunities within digital humanities is the ability to bring together methods, approaches and topics from a range of disciplines and use them to conduct genuinely interdisciplinary research. In this chapter we bring together methods from corpus linguistics and geographical information science to conduct a piece of analysis relevant to both historians and geographers. We provide a case study of how

UK place-names are used in relation to the term *poverty* in *The Times* newspaper between 1940 and 2010. We use a new technique, geographical text analysis (GTA) (Donaldson et al. 2017; Gregory and Donaldson 2016; Porter et al. 2015), to demonstrate how corpora – and particularly the corpus linguistics tool of concordance – can be used in combination with methods from geographical information science (GISc) to foreground, visualise and analyse explicit mentions of place within texts. (For good introductions to GISc and the closely related field of geographical information systems [GIS] see, for example: Chrisman 2002; Heywood et al. 2006; Longley et al. 2015.) As such, we present an innovative approach to the study of long-term geographical change from very large digital collections that could be applied in many other contexts.

Background: the use of newspaper corpora

Some of the largest corpora available to historians and the academic community more broadly come from newspapers. Part I of the British Library's 19th Century Newspaper collection on its own contains, where available, continuous runs for the whole century of 47 newspapers from across Britain and Ireland (see: www.gale.com/uk/c/british-library-newspapers-part-i). It consists of almost 2 million pages and at least 30 billion words of text. Historians have realised the potential value of these sources and accepted the need for new approaches to their study (Mussell 2012; Nicholson 2010). These approaches particularly focus on the use of corpus linguistics to allow a combination of quantitative and qualitative explorations of the language used within the text (Adolphs 2006; McEnery and Hardie 2012). A simple example of this is provided by Nicholson (2012) who explores the changing frequency of words such as 'America' and 'Germany' and other words that they are associated with these countries using the British Library 19th Century Newspaper collection. These approaches have also been adopted by scholars exploring modern newspapers to research, for example, themes associated with minority groups (Baker et al. 2014), immigration (Gabrielatos and Baker 2008) and political correctness (Johnson et al. 2003).

There are two recurring issues in the use of historical newspapers within historical research. From an applied perspective, all newspapers have political, editorial and other biases which may change over time (Brake and Demoor 2009). The way that they represent the word is thus not impartial. Instead, exploring the biases in the themes and places that newspapers cover, and how these biases were represented, becomes an area of active research interest (Hardie and McEnery 2009). The second, more pragmatic one, is the issue of optical character recognition (OCR) errors. Bulk digitisation of texts is most cost-effectively conducted by first scanning the page and then using OCR software to convert the pixels on the image into the alpha-numeric characters that a computer can understand as text. The difficulty is that, although quick, OCR is an error-prone process, and a significant number of words are likely to have one or more characters that have been incorrectly identified by the software. This, in turn, means that a computer will not recognise the word in, for example, a keyword search. Historic newspapers are particularly prone to errors of this sort, as the quality of both the paper and the printing is likely to be relatively poor (Holley 2009). Using the British Library 19th Century Newspaper Collection, Tanner et al. (2009) estimated that only 83.6 percent of characters were accurate leading to a word accuracy of only 78.0 percent which declined to 63.4 percent for words that started with a capital letter. The impact of these errors has led to significant criticism (Hitchcock 2013; Putman 2016). However, detailed research suggests that, from a corpus linguistics perspective, poor OCR may not be as significant an issue as has been suggested (Joulain-Jay 2017).

In this study, we explore *The Times* from the 1940s to the 2000s. *The Times* is widely regarded as Britain's newspaper of record. It has been published since 1785 making it the longest continuously published newspaper in the world. It has always been published in London and is generally regarded as being aimed at an educated, middle-class market. The newspaper has been digitised for the period from 1785 to 2012 as *The Times Digital Archive* (see: www.gale.com/uk/c/the-times-digital-archive). We have access to the underlying text, hosted in CQPweb (Hardie 2012) at Lancaster University, under licence from Gale Cengage.

Issues and research in the study of poverty

The study of the geographies associated with poverty has a long history stretching back at least to Charles Booth's maps of poverty in Victorian London (see: <https://booth.lse.ac.uk>). Much geographical research on poverty has been quantitative, based on identifying poverty and inequality from sources such as the census or household surveys (Dorling 2013; Thomas and Dorling 2011; Townsend 1979). There have also been a number of studies which have explored changing geographical patterns of poverty over time in London (Dorling et al. 2000), England and Wales (Gregory 2009) and Britain as a whole (Dorling 1997). Working at different scales, these studies have all shown that patterns of poverty are very strongly entrenched. With very few exceptions, poor areas remain poor relative to other places, while rich ones remain relatively rich. These studies are all based on quantitative representations of how unequal or impoverished areas are. Poverty is, however, a notoriously difficult concept to define, and quantitative measures, such as unemployment, low-skilled work, and overcrowded housing, are well known to be surrogates for what is a complex and multi-faceted phenomenon.

In this chapter, we are interested not in poverty as defined quantitatively, but instead as it is represented in textual sources, namely newspaper corpora. Our particular interest is in whether the temporal stability which geographical studies have found in quantitative poverty are also found in newspaper representations of impoverished places. In other words, we want to find out which places *The Times* newspaper associates with poverty and investigate whether these places remain constant over time. To this end, we trace the relationship between the use of place-names and the term *poverty* in *The Times* from the 1940s to the start of the twenty-first century. While we have chosen to demonstrate our method by focusing on a single corpus query – we use `<*poverty*>`¹ to locate terms such as *poverty*, *anti-poverty*, *poverty-stricken*, etc. – larger-scale and more robust GTA depends on the analysis of a larger number of search terms to build a picture of how place is used discursively in relation to a given topic. This large-scale GTA is exemplified by Paterson and Gregory's (2019) analysis of discourses of poverty and place in the *Guardian* and *Daily Mail* newspapers between 2010 and 2015.

For the present chapter, we use seven decades from *The Times*: the 1940s, 1950s, 1960s, 1970s, 1980s, 1990s and 2000s. In interrogating how place is used during different periods in British history, we can trace how notions of poverty move in space as well as in ideological, social, political and cultural realms. For example, our 1940s corpus covers the end of the Second World War, rationing and austerity, as well as the founding of the National Health Service (NHS) and the beginnings of the welfare state. In 1950 Joseph Rowntree published his third report on poverty, which suggested that post-war poverty rates had declined dramatically, although reanalysis has indicated that this was an overstatement of the case (c.f. Piachaud and Webb 2004: 45). At the end of the 1950s, in 1957, Prime Minister Harold Macmillan claimed that most Britons had 'never had it so good',

Table 22.1 Hits for <*poverty*> in *The Times* corpora

	Hits	Freq. pmw	No. of texts	Avg. hits per text	Total corpus word count	Total corpus texts
1940s	1,768	6.09	1,153	1.53	290,448,326	3,100
1950s	2,230	5.30	1,337	1.66	420,978,622	3,087
1960s	4,359	9.39	1,956	2.23	519,523,147	3,098
1970s	6,129	12.75	2,095	2.93	480,659,269	2,789
1980s	6,651	12.38	2,373	2.80	537,031,328	3,090
1990s	8,614	11.23	2,610	3.30	767,266,910	3,125
2000s	14,500	17.63	2,913	4.98	822,524,677	3,122

but the apparent prosperity of the decade was followed by wage freezes in the 1960s and rising inflation rates. Piachaud and Webb (2004: 50) note that poverty was ‘rediscovered in the 1960s’, a decade which included the founding of the Child Poverty Action Group. Poverty rates decreased in the 1970s, but by the end of the decade they were set to rise dramatically. The 1980s corpus represents *The Times*’ response to Thatcherism and deindustrialisation, a period of time when poverty rates continued to rise. The first *Breadline Britain* survey was conducted in this decade (Mack and Lansley 1985), which estimated that 7.5 million people in Britain were living in relative poverty (1985: 182). There was a recession in the early 1990s, but in the years that followed there was decreasing unemployment and the introduction of the minimum wage in 1999. Finally, the 2000s corpus includes initial economic prosperity followed by the downturn of the global economic crash in 2008.

All of these historical events and measurements have a relationship to poverty in both its economic and social conceptualisations. As a crude measure of media interest in poverty we can compare how often the term is used across our seven corpora (see Table 22.1). Mentions of poverty decline from a normalised frequency of 6.09 tokens per million words (pmw) in the 1940s to 5.30 pmw in the 1950s. There is an increase in mentions throughout the 1960s and 1970s with a plateau in the 1980s corpus. Mentions of poverty decline in the 1990s before reaching their peak in the 2000s corpus. However, far from prompting conclusions about poverty in Britain during this time span, these figures do not provide information about what *The Times* was saying about poverty in each of these decades, nor whether the poverty written about was geographically located. The 1980s corpus, for example, contains reports about the Ethiopian famine alongside references to London and Philadelphia – the locations of the Live Aid charity concerts held to raise funds for famine relief in 1985. Thus, the poverty referenced in these cases is not located within the UK despite the use of the UK place-name ‘London’.

In order to understand which locations *The Times* associated with poverty (in the UK or elsewhere) we can turn to GTA. What GTA can tell us is how, or indeed whether, place was significant in *The Times*’ representations of poverty during each of our chosen decades. Specifically, we pose the following research questions:

- 1 Which UK places are associated with poverty by *The Times*, and are these associations stable over time?
- 2 Why are particular places associated with poverty?
- 3 How is the relationship between poverty and place characterised in each decade?

Research methods: geographical text analysis

GTA involves combining methods from corpus linguistics and GISc. For the uninitiated, corpus linguistics involves the computationally aided analysis of large bodies of texts. It facilitates the analysis of language at a scale beyond that possible using manual analysis alone and many currently available corpora, especially those which have been constructed using web crawling technology, contain billions of words. The corpora we are using here contain a combined total of 3,838,432,279 words split across seven decades (Table 22.1). Corpus linguistics tools include the calculation of word frequency lists, keywords (those words which occur at a statistically higher rate in a given corpus compared to a reference corpus), collocations (the relationship between words) and concordances (all examples of a given query term in its immediate co-text) (see, for example, Adolphs 2006; McEnery and Hardie 2012). GTA draws particularly on concordances – the case study presented here uses a form of geoparsing known as *concordance geoparsing* – but this is not the only corpus tool that can be useful for GTA. As an example, Paterson and Gregory (2019) used collocates and contrastive collocation analysis (see later) to illuminate trends in discourses of poverty which illustrated traces of underlying ideologies, such as flawed consumerism (Bauman 2005) and the deserving/undeserving poor (Katz 2013).

The tools of corpus analysis are often used in triangulation with other methods of linguistic analysis. For example, there is an established tradition of combining corpus analysis with discourse analysis (see Baker 2006). Increasingly, too, corpora are being used beyond the realms of linguistics within the wider social sciences and the humanities (see <http://cass.lancs.ac.uk/> for examples of such initiatives). By drawing on geographical methods of analysis, this chapter is an example of the expansion of corpus linguistics tools beyond the boundaries of the discipline of linguistics. In particular, we combine corpus linguistics with methods from GISc to implement a GTA. This requires us first to convert the unstructured text from a corpus into a structure that can be used within GIS and then to use techniques from both GISc and corpus linguistics to explore the geographies found within the corpus.

A major challenge for GTA is thus how to convert unstructured text into the tabular form that GIS software requires. This is implemented by making what we term the place-name co-occurrence (PNC) the basic unit of analysis within GTA. PNCs consist of the query term, the place-name, the co-text that surrounds them and the point location (the geographical coordinates) that can be used to map that place-name. As shown in Figure 22.1, it is possible for a single concordance line to produce multiple PNCs; there are two occurrences of ‘London’ within the 15-word span of the node ‘poverty’, and so these represent two separate PNCs. For examples of studies in the spatial humanities using the concept of PNCs see Gregory and Donaldson (2016) and Donaldson et al. (2017). Analysis using PNCs assumes that co-occurrence of a query node and a place-name is indicative of a relationship between the two. At its most basic, this relationship can be expressed in the Firthian sense that ‘you shall know a word by the company it keeps’ (Firth 1957: 11). Murrieta-Flores et al. (2015: 300) note that proximity of place-name mentions to a query node does not guarantee this link; nevertheless, they come to the conclusion that work on collocations ‘seems to suggest that simple proximity, with no explicit connection, if repeated consistently, is indeed enough to create an implicit link between two meanings in the mind of speakers and hearers’.

Creating PNCs from large corpora requires a tool known as a geoparser. A geoparser first identifies likely place-names using named entity recognition (NER) techniques and then uses a gazetteer to assign geographical coordinates to all occurrences of place-names (Grover et al. 2010). Place-names are, however, complicated things that are difficult to

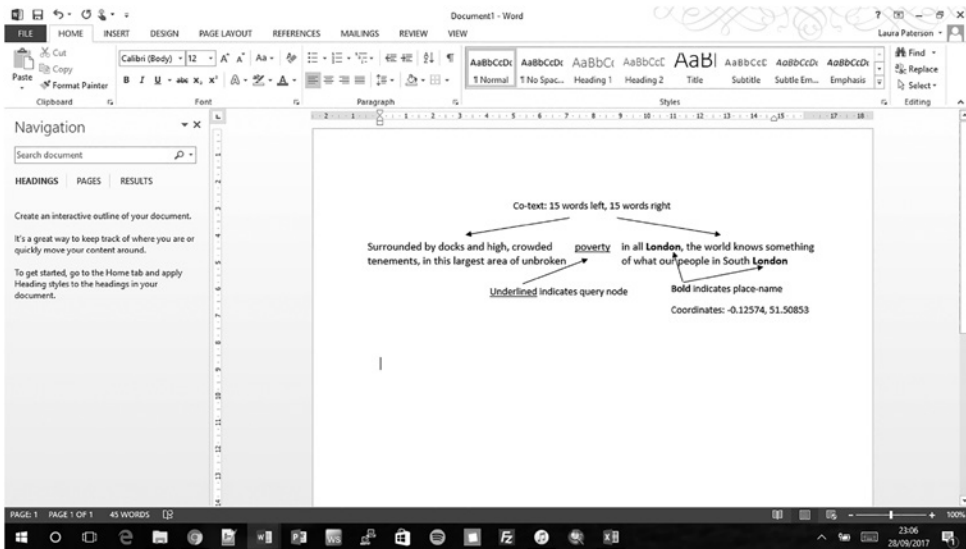


Figure 22.1 Example PNC from *The Times* 1940s corpus

process in an entirely automated manner. To raise standards of accuracy in both identifying place-names and in allocating them to an appropriate co-ordinate, we use a process called concordance geoparsing in which only the immediate co-text around the search-term is geoparsed, allowing for easier error checking and correcting (Rupp et al. 2014). The nature of concordance geoparsing is such that it effectively returns PNCs as the results include the search-term, place-name, co-ordinate and surrounding co-text. One critical choice in defining a PNC is setting the maximum span between the search term and the place-name for the concordance to be included. In this case, we chose a span of 15 words/tokens (punctuation counts as one token) either side of the query node. While we have used a smaller span of ten words elsewhere (Paterson and Gregory 2019), we chose to increase the span here to account for the fact that there were some OCR errors in the 1940s corpora as they had been scanned and converted to a digital format.²

Figure 22.2 shows five examples of concordance lines that will generate PNCs using the ± 15 token span. Once the PNCs are generated, they are downloaded as a plain-text file (including their metadata, such as their text ID, date of publication, etc.). This process can be repeated for multiple queries, although here we focus on just one query (*<*poverty*>*) to demonstrate the method. The first concordance line in Figure 22.2 generated two PNCs (one for Liverpool and one for Broadstairs). By contrast, the final example in Figure 22.2 generates one PNC (London) as we did not analyse the texts down to street level; thus ‘London Bridge Street’ as not considered to be its own PNC. Once the PNCs have been generated, the resulting files were then opened in flat-file database software (in this case Microsoft Excel) to be cleaned.

Cleaning involved, for example, removing erroneous hits from our results (i.e. ‘hull’ referring to the hull of a ship as opposed to Kingston-Upon-Hull in Yorkshire). It may also involve disambiguating between different places with the same name, such as ‘Newcastle’, which could refer to Newcastle-upon-Tyne or Newcastle-under-Lyme. We also removed references to poverty that comprised part of TV and/or radio listings, as well as more figurative

the film he made 20 years ago when he determined to right the wrongs of	poverty-stricken	Liverpool, ' and compares his life then to his life today in Broadstairs 140
, if he came to that conclusion, in the next session. "	poverty	Bill introduced Mr Jeremy Corbyn (Islington North, Lab) was given leave
cent higher than the figure for Glasgow as a whole. " Quite simply,	poverty	kills, " he said. Deaths in the first year of life in Easterhouse
". Many of the buildings on the London list are there not through inherent	poverty	but are victims of an unsympathetic owner or a slow – moving planning machine.
safe and healthy place to live and work; and a SFK determined push on	anti-poverty	strategies and 32 London Bridge Street community development that will foster an inclusive London SE

Figure 22.2 Example concordance lines for <*poverty*> in *The Times* 1990s corpus

uses of poverty, such as a ‘poverty of titles’ which referred to not having suitable music to play, and the description of low-quality rugby (‘poverty play’) and cricket (‘poverty of wickets’). We were concerned with the relationship between poverty and place in the UK only, so we removed references to other countries, such as America and France. Additionally, we removed country-level references, such as Scotland, Wales, etc., in order to focus more closely on local areas within the UK. Once the data were clean, we uploaded the final database files into database management and cartographic software tool ArcGIS to analyse, and ultimately visualise, how poverty and place are used within *The Times* newspaper from 1940 to 2010.

Sample analysis: the geography of poverty in *The Times* 1940–2010

Table 22.2 shows the distribution of hits for the query <*poverty*> (henceforth expressed as *poverty*) by decade. Column 2 of the table suggests that *The Times*’ interest in poverty more than doubled between the 1950s, when there were 5.3 hits pmw, and the 1970s with 12.4 hits pmw. Since the 1970s, however, the interest has remained roughly constant. Column 4 shows that overall there is a link between poverty and place. It shows that over the entire period, for every thousand instances of *poverty*, 28.7 of them had a place-name within 15 tokens. In other words, poverty is associated with place to a certain extent. Interestingly, unlike the overall instances of *poverty* in the corpora, there does not seem to be a clear pattern of change over time in the link between poverty and place. This suggests that whatever caused interest in poverty to rise over time, it was not directly linked to geography, at least not geography as revealed by the explicit use of place-names. Column 5 shows that when normalised by the overall size of the relevant corpus, interest in poverty and place, as revealed by *poverty* PNCs, rose through the 1950s and 1960s to peak in the 1980s before dropping back. Poverty was thus far more closely associated with place in the 1980s than in other decades. The 1940s are also interesting, but also exceptional; although there was a relatively low interest in poverty overall in the 1940s, it appears to be far more associated with geography than any other decade.

Table 22.2 Instances of <*poverty*> in *The Times* corpus by decade (pmw)

	(1)	(2)	(3)	(4)	(5)
	Query hits	Freq. per million	PNC poverty +/- 15	Normalised PNCs per query hits (per thousand hits)	Normalised PNCs pmw (whole corpus)
1940s	1,768	6.09	95	53.73	0.33
1950s	2, 230	5.30	73	32.74	0.17
1960s	4,359	9.39	121	27.76	0.23
1970s	5,952	12.38	157	26.38	0.33
1980s	5,977	11.13	242	40.49	0.45
1990s	8,614	11.23	215	24.96	0.28
2000s	10,408	12.65	232	22.29	0.28
TOTAL	39,308	10.15	1135	28.74	0.29

This does not, of course, tell us anything about where *The Times* associates with poverty, but rather that it links poverty and place to a degree that varies over time. However, as the geoparsing process provides us with a coordinate for each place-name within a PNC, we can conduct spatial analysis to explore the geographies of poverty and place. The coordinates form spatial data that allow us to read the PNC into GIS. The easy way to map these data would then be to use a dot map where each coordinate is represented by a point.

Figure 22.3 shows the pattern of PNCs in the 1940s and the 1980s. It shows that in the 1940s poverty instances are only found in London and Newcastle, indeed almost all of these are in London. By the 1980s the pattern is much more dispersed, particularly across England. This type of mapping is not, however, a particularly effective solution for visualising the PNCs. Dot maps are difficult for the human eye to interpret, particularly as large numbers of points may be at exactly the same location and thus disappear under each other, as in Figure 22.3a in particular. For this reason, Figure 22.4, which shows all of the *poverty* PNCs in *The Times* corpora, also uses a process called density smoothing to simplify and summarise the pattern. Density smoothing effectively measures the number of point locations, in this case PNCs, which are near to each location on the study area (Lloyd 2007). The more PNCs there are near the location, and the nearer they are, the higher the value at that location. A critical factor in implementing this is how we define ‘near’. This is done using a bandwidth that means that the weighting given to each PNC declines with its distance from the location. The choice of bandwidth can be termed quantitatively based on the distribution of points around the study area (Fotheringham et al. 2000: 149). To be consistent, and allow comparisons between decades, we have normalised all of the densities using z-scores. A z-score of 0.0 shows the mean density, 1.0 is one standard deviation above the mean, and so on. This allows us to identify clusters based on the idea that values with a z-score of above 1.96 are significant at $p < 0.05$ and those above 2.58 are significant at $p < 0.01$ although, for various statistical reasons, these thresholds should be seen as exploratory rather than as formal measures of statistical significance.

Figure 22.4 thus highlights clearly that *The Times* strongly associates poverty with London; the only place on the map that lies above the $z > 1.96$ threshold. Of the 1,131 PNCs across all corpora, 53.1 percent are within Greater London, suggesting that *The Times* is far more focused on poverty in London than anywhere else. This figure, in itself, highlights one of the advantages of the spatial analytic approach over a conventional corpus linguistic approach:



Figure 22.3 Dot maps showing the pattern of PNCs in the 1940s (left) and 1980s (right)

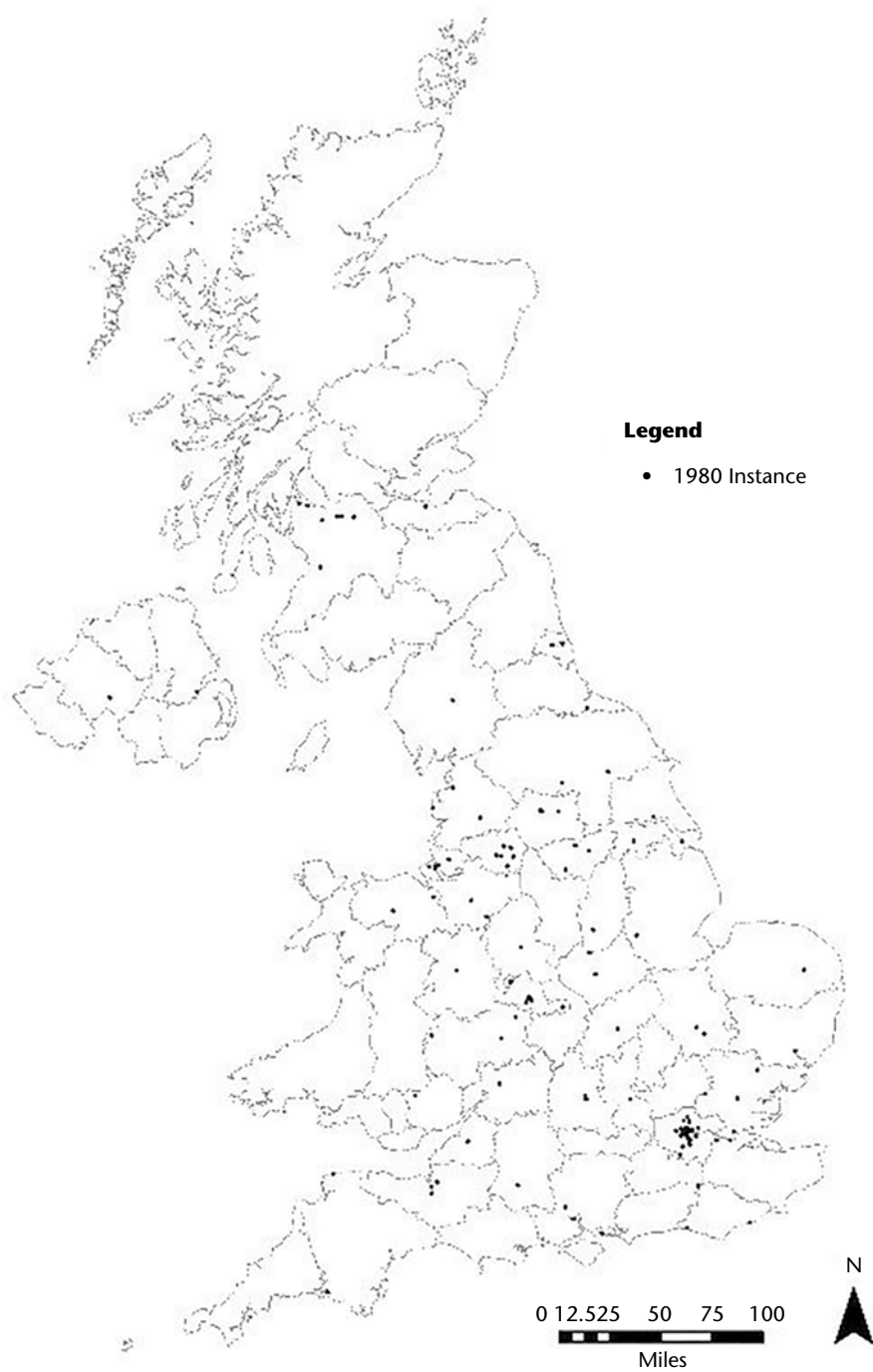


Figure 22.3 (Continued)

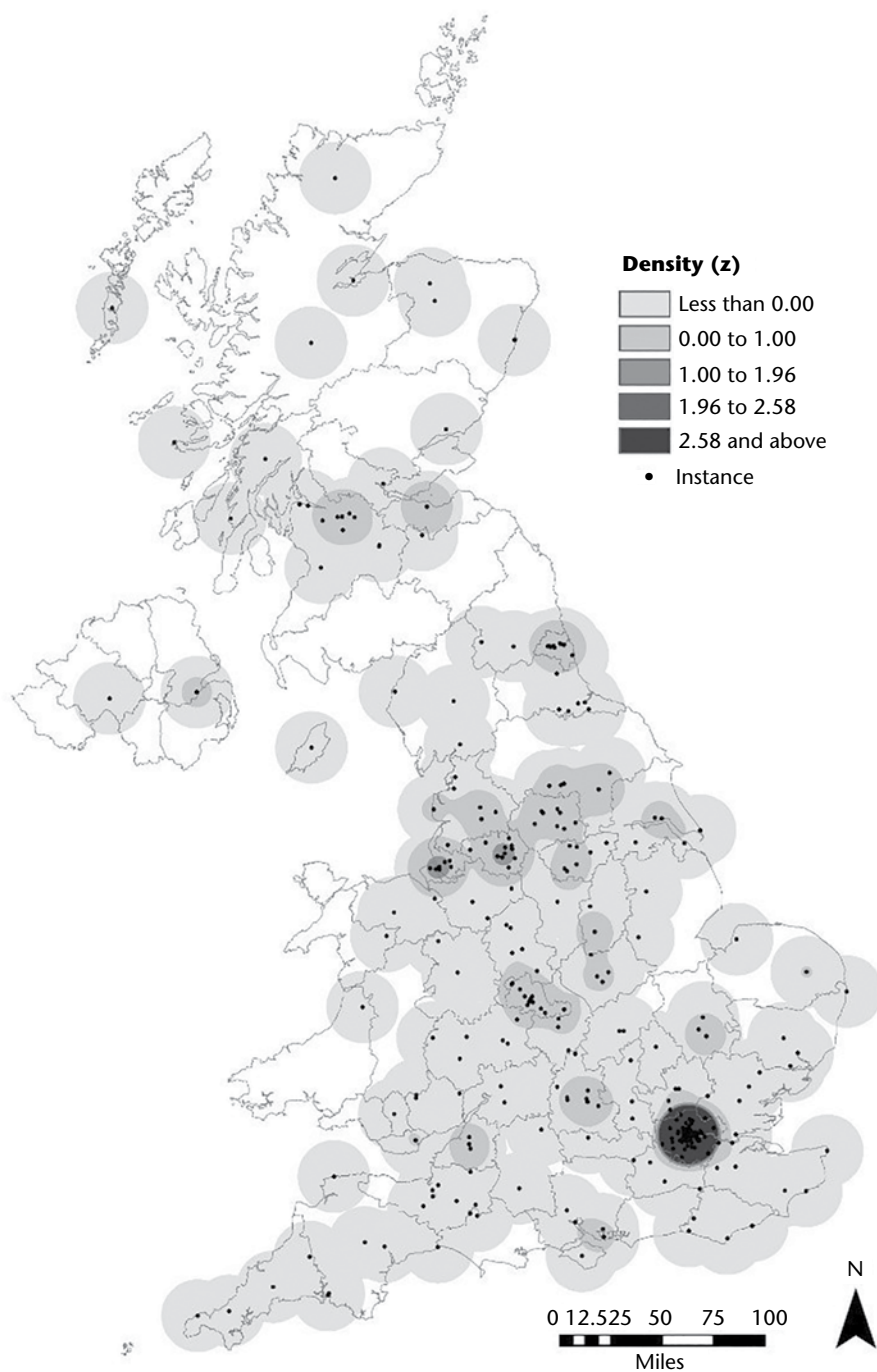


Figure 22.4 Density smoothed map of all <poverty> PNCs from *The Times*, 1940–2009

while corpus linguistics would be able to show that there were more co-occurrences between, for example *poverty* and ‘London’, than other place-names, it would be unable to automatically include named places within and around London, such as Westminster, Tower Hamlets, and Islington within this calculation. Liverpool and Manchester, with 41 and 40 PNCs, respectively, have smaller clusters and there is also raised interest in poverty in many other urban places such as Belfast, Glasgow, Edinburgh, Newcastle and so on.

It would be possible to produce maps similar to Figure 22.4 for each decade; however, the similarities and differences between these maps are difficult to present in paper form, in particular, to compare over time. Instead, Figure 22.5 divides the country into three major macro-areas: first, Greater London; second, the South which includes all counties outside of Greater London that are south of the arbitrary line running from the Severn Estuary in the west to the Wash in the east, which is extensively used to define the UK’s north-south divide; and third, the North: all counties north of the Severn-Wash line including those in Scotland, Wales and Northern Ireland. The North is then further sub-divided into places in the historical counties of Lancashire and Yorkshire – which include many major industrial cities including Liverpool, Manchester, Leeds and Sheffield – and ‘Other North’ which is the rest of the UK north of the Severn-Wash line.

Figure 22.5 does not include the 1940s as this decade is a strange anomaly – all but one of its 95 PNCs are in London. For other decades, however, the coverage of poverty is dominated by either London and/or the North, while the South is consistently not associated with



Figure 22.5 Macro-geographies of <*poverty*> instances over time, 1950s onward

poverty. For most decades, London attracts more coverage than the North, typically over 50 percent of the total PNCs. The exceptions to this are the 1950s and – marginally – the 1980s. In the 1950s this exception is driven by a spike in interest in poverty in Lancashire and Yorkshire. The 1980s appear to be a pivotal decade in interest in poverty in the North: PNCs in Lancashire and Yorkshire rise from being around 10 percent of the total to being around 20 percent. At the same time, interest in poverty elsewhere, in the North, which had been rising steadily since the 1950s, peaks at 24.4 percent of PNCs, from where it declines in subsequent decades. This combination accounts for instances of *poverty* in the North narrowly exceeding interest in poverty in London in the 1980s.

This type of spatial analysis is very good at identifying patterns. However, these patterns pose questions to the researcher to then try to explain them. The major questions that emerge from these findings are: Why is poverty only associated with London in the 1940s? What is it overall about poverty in London after the 1940s that attracts attention? How does what is said about London compare with other parts of the country? What is it about Lancashire and Yorkshire, in particular, in the 1950s that attracts attention? And, why does interest in Lancashire and Yorkshire increase in the 1980s?

To begin to answer these questions we need to move away from GIS and back to corpus linguistics. There are two approaches to answering these questions: the first is qualitative and simply involves exploring the corpus to perform close analysis of the concordance lines in which the place-names co-occur with the query node within the set span of ± 15 words. The second is more quantitative and involves contrasting the co-text of one set of PNCs with a second set.

When there are relatively small numbers of PNCs, the qualitative approach is usually sufficient on its own. To understand what it was about poverty in London in the 1940s we needed to explore 94 PNCs. Eighty-two of these PNCs refer to 'London' itself; the others are places such as Whitechapel, Stepney, Bethnal Green and Westminster. Exploring their concordance lines reveals that the majority are concerned with charity appeals during the war to help people in impoverished parts of London. The concentration of these is perhaps somewhat misleading as some of these appeals are repeated multiple times. For example, an appeal by the South London Mission which included example (1) was repeated 18 times through November and December 1940.

- (1) Surrounded by docks and high tenements, in this largest area of unbroken *poverty* in all **London**, the world knows something of what our people in South London are enduring in the Battle for Freedom. Surely this year, in brave hearted South London Father Christmas is needed more than ever. Your Christmas gift will help them hold on until Victory

The other interesting thing is that only nine of the PNCs are in the post-war period, and these are again largely associated with charitable appeals. This shows that in the period while the welfare state was being created, there was little, if any, consideration in *The Times* of the geographical aspects of poverty, at least as revealed by our single search-term. Exploring the negative – why does *The Times* ignore poverty everywhere else in the country in the 1940s? – is, of course, more difficult.

When there are larger numbers of PNCs, a more quantitative approach is required, as simply looking at concordance lines is insufficient. To explore the interest in poverty in London in comparison to other places from the 1950s onwards means exploring 507 PNCs in London and 529 elsewhere. The question thus becomes, what is it about PNCs in London that is different for those from elsewhere? To explore the similarities and differences between how poverty is related to place in different geographical locations, we use a specialist type

Table 22.3 PNC keywords derived from a comparison of the PNCs for Greater London with those for elsewhere from the 1950s onwards. Ranked by log-likelihood ($p < .01$, *Italics* $p < .05$. Observed ≥ 5). Punctuation, place-names, and numbers excluded.

Cluster	PNC Keywords
London	group, child, Victorian, I, war, action, want, against, tiny, genteel, wishes, deduction, by, homelessness, big, suburbs, price, press, greatest, Peabody, evidence, administer, deal, <i>society, link, hall, an, supplementary, rights, some, familiar, may, little, old, much</i>
Not London	lab, Mr, sir, industrial, September, shopping, had, said, incomes, this, would, MP, rather, grown, effects, escape, measure, imagination, <i>trap, when, times, been, primary, another, planning, asked, town, first, left, right, into, church, estate, not</i>

of contrastive concordance analysis. To do this we compared the co-text element of the PNCs for London with the co-text element of the PNCs for the rest of the country in order to identify what words are associated more with one set of PNCs than the other. We call the resulting sets of words PNC keywords.

In Table 22.3, log-likelihood, one of a range of possible statistics, has been used to identify statistically significant PNC keywords for poverty from the 1950s onward. Close analysis of the London PNC keywords reveals that a lot of poverty charities are based on London ('group', 'child', 'action', 'against', 'Peabody' and some instances of 'war'). There is also some discussion of, or comparisons with, poverty in the past or in history ('Victorian', some instances of 'war' in senses such as 'since the war' or 'before the war'). 'Greatest' is often used in the sense of a comparison with London having both the greatest affluence and the worst poverty (2).

- (2) . . . while **London** has both the greatest affluence and also some of the worst pockets of poverty. The North West and **London** have most districts falling within the country's 150 most deprived districts.
- (3) A mention in your will can bring food and shelter to tiny hungry children and to old people. – WAR ON WANT Campaign Against World Poverty. 9 Madeley Road. London.

Another interesting adjective is 'genteel', as 'genteel poverty' seems to have been a condition frequently referred to in London. Beyond this, 'homelessness' is a key theme and poverty in London also seems to be associated with 'old' people (3). Attempts to alleviate poverty are discussed in the sense of 'Supplementary' benefit, 'evidence' of poverty and the need to 'deal' with poverty, and there is also 'press' attention.

Beyond the capital, politics seems much more important, particularly in discussions of debates in parliament where politicians representing non-London constituencies are frequently quoted ('Lab' [in the sense of the Labour party], 'MP', 'Mr', some instances of 'Sir'). There is also considerable discussion about poverty in the letters pages. Letters either discuss poverty occurring outside London or have been written by people outside London (other instances of 'Sir', 'Times', 'September'). In example (4) we have two PNCs generated because the letter writer lived in Houghton (Hampshire) but was making a derogatory comment about people from Devon.

- (4) Sir, On a recent visit to the library in a small market town in **Devon** I found a large volume entitled Poverty in the careers section. Yours faithfully, SUSAN ELKS, The Willows, **Houghton**

Using examples like this we can also begin to see the types of stance taken in *The Times* towards poverty; although (4) is not an example of text authored under the newspaper's masthead by an in-house journalist, the letter was selected for print by an editorial gate-keeper. Although it is not sufficient to assume that this automatically implies *The Times*' endorsement of the sentiment expressed in (4), we can at least take this as an example of the types of letters concerning poverty that *The Times* saw fit to print.

There are also more PNC keywords directly associated with poverty than are found in London, for example 'grown' up in poverty in a place, the 'effects' of poverty or unemployment in a place, the poverty 'trap', and 'primary' poverty ('primary' is also sometimes used in the context of primary schools). Housing reveals itself in a different way to London, as elsewhere poverty is associated with 'estate(s)' and also with 'town(s)'. Poverty outside London is also associated with 'industrial' areas or sectors (5). There is little on what can be done about poverty except in some aspects of 'planning' and in actions by the 'church.'

- (5) Textile mills cling to the sides of the **Calder** valley, providing an industrial base on which the prosperity or the poverty of the area depends. Rugged, wind-swept hillsides provide a foot-hold for small farms whose occupants, more out of tradition than anything else, tend to lean towards the political right

Our third question concerned what it was about poverty in Lancashire and Yorkshire that attracted attention in the 1950s. Here qualitative analysis is again sufficient as there are only 73 poverty PNCs for the country as whole in the 1950s. Of these, 22 were in Lancashire and Yorkshire, almost all of which – 17 – were in York. The main reason for this is that 1951 saw the publication of Seebom Rowntree's final report on poverty in York *Poverty and the Welfare State* (Rowntree and Lavers 1951) (the earlier ones had been published in 1901 and 1936) which, as well as attracting interest in its own right, also seems to have made York the example place for poverty in this decade.

The 1980's interest in Lancashire and Yorkshire is more diverse both in terms of geography and theme. There are 42 PNCs in this region over the decade, and they cluster around Liverpool, Manchester, Bradford and Humberside in particular. The co-text of these PNCs is mainly concerned with levels of poverty in these places, sometimes in relation to the riots that occurred in the early 1980s, especially in Liverpool, with links between poverty and race being another prominent issue that emerges from the 1960s onwards. There are also strong hints about the impact of industrial decline in these areas although poverty is frequently associated with the industrial era as well as the 1980s. An example of this is:

- (6) . . . when Britain was the workshop of the world and when the ports were full of ships and ship-building, cities like **London, Glasgow, Liverpool, Manchester** and **Newcastle** were notorious for their poverty, their overcrowded slums, their terrifying juvenile mortality, crime, prostitution, disease and destitution, as well as for the existence of a savage urban underclass.

It is clear from this that the geographies of poverty reporting undergo major changes over time, particularly with an increased focus on Lancashire and Yorkshire and a decline in interest elsewhere in the north over the same period. To explore this in more detail we needed to both lower the scale of the investigation to remove, for example, the impact of a single city, such as York in the 1950s, having such a marked influence on broader regions, and to analyse whether these changes are statistically significant. Doing this involves returning

to spatial analysis where we can use a technique called spatial segregation analysis which addresses both issues (Diggle et al. 2005; Rowlingson 2015). Spatial segregation analysis involves comparing two point patterns, thus avoiding aggregation, to see whether they appear to have statistically significantly different distributions and, if so, to identify where on the map they are different.

Spatial segregation analysis was used in two different analyses: the first compared the 1980s with other decades to test whether the 1980s appeared to be exceptional, the second compared the period before the 1980s with the 1980s and later decades to see whether the 1980s appeared to represent a shift in the pattern. In both cases the 1940s were excluded as they have an exceptional geography.

The overall results comparing the 1980s PNCs with other decades shows that the two patterns are not significantly different ($p = 0.11$). Figure 22.6, however, explores the more local patterns identifying the contour lines where the PNCs in an area are significantly more likely to come either from the 1980s or from other decades. It shows that the northwest of England was significantly more associated with poverty in the 1980s than it was in other decades, while the southeast had significantly less in the 1980s, although only at the $p < 0.05$ level. Interestingly, the $p = 0.50$ contour, the level at which there is an even possibility of the PNC being from either decade, runs almost perfectly along the Severn-Wash line. This suggests that the 1980s shows an increased north–south divide in *The Times*' reporting of poverty compared to other decades. The more northerly $p = 0.50$ contours, such as the one that runs through northern and eastern Scotland, can safely be discounted as they are based on very small numbers.

Figure 22.7 repeats the segregation analysis, but this time compares the decades from the 1950s to the 1970s with the 1980s and later. In this case there is a statistically significant difference between the two sets of PNCs at $0 < 0.01$. The local pattern, shown in Figure 22.7, is in many ways similar to the pattern shown in Figure 22.6. The northwest of England became significantly more associated with poverty after 1980, while areas that become less associated with poverty in the later decades are again in the south and east, albeit this line is now somewhat farther north than in Figure 22.6. The area where interest in poverty has dropped the most, in other words where it was higher in the pre-1980s period, was the Midlands, while there was no statistically significant change in the number of PNCs in London. Overall, these two maps do suggest that the 1980s was a pivotal decade when *The Times*' interest in poverty shifted towards the northwest of England and away from the south and, in particular, the Midlands.

Conclusions

Early examples of GTA using sizeable corpora, of which this chapter is one example, demonstrate the fruitfulness in considering place in an analysis of language. Simultaneously, they show how analysis of textual data can inform analysis of geographical patterns. The preceding sections have demonstrated that not only does the relationship between poverty and place fluctuate temporally in *The Times*, particularly in relation to the wider social context, economic trends and political policies, but that different parts of the UK are associated with poverty in different ways. London, although it dominates the PNCs, seems to be the location of anti-poverty charities (see the data from the 1940s, and references to the Child Poverty Action Group), whereas northern cities, tend to be associated with the relationship between poverty and industrial decline. Poverty in the north also appears to be constructed as a state of being: people grow up in poverty, are in the poverty 'trap', etc. We have drawn

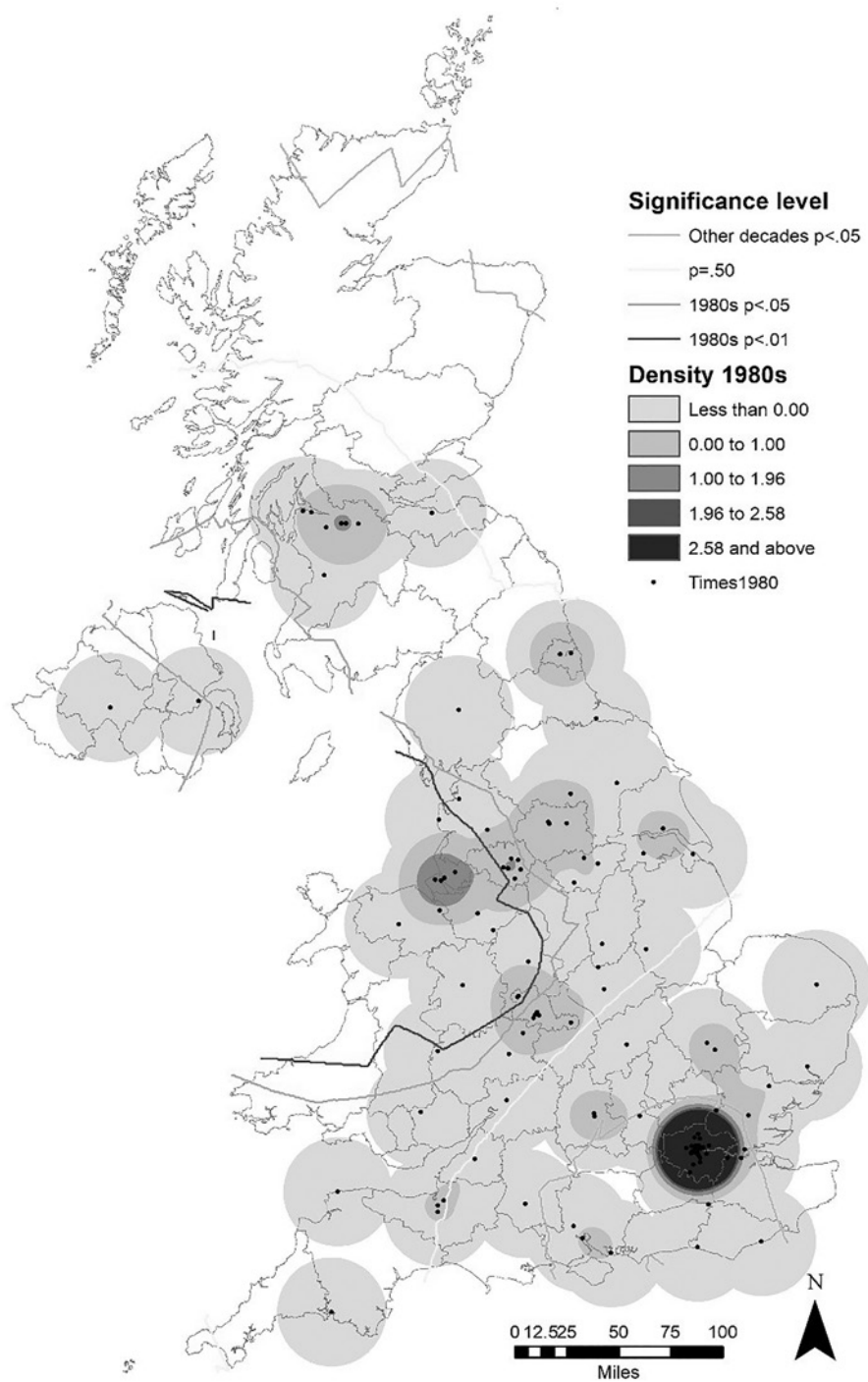


Figure 22.6 Spatial segregation analysis comparing the locations of PNCs in the 1980s with those from other decades from the 1950s on. PNCs and densities are for the 1980s

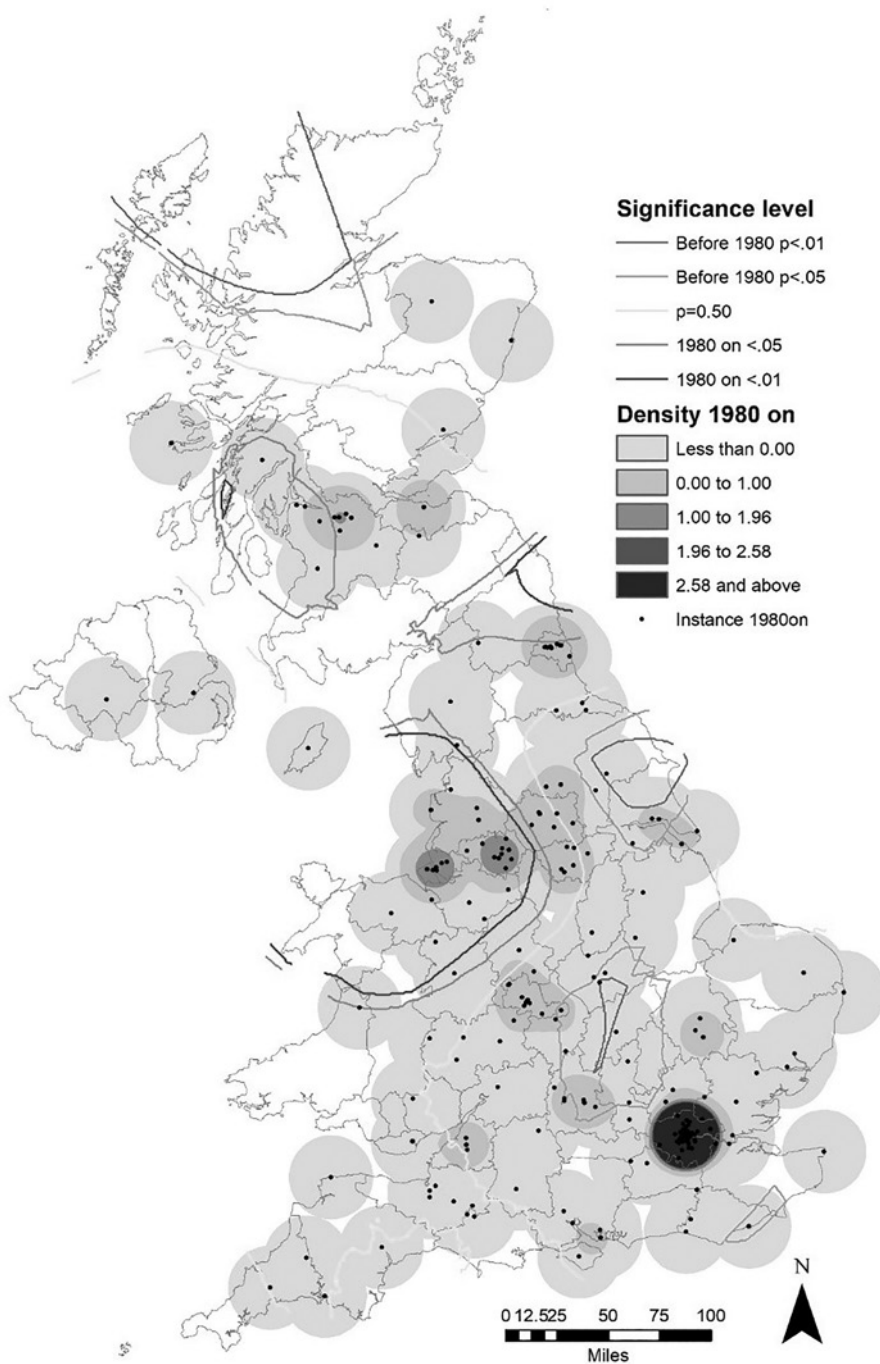


Figure 22.7 Spatial segregation analysis comparing the locations of PNCs from the 1980s onwards with those from the 1950s to 1970s. PNCs and densities are for the 1980s onwards

these conclusions based on a single search term, but rather than this being a limiting factor, we argue that now these relationships between poverty and place are apparent, they can form the basis of hypotheses about the geography of *The Times*' coverage of poverty.

We acknowledge that there are limitations to the geoparsing process; while the concordance geoparser can account for certain spelling variations of place-names, the generation of PNCs is, nevertheless, restricted to explicit mentions of identifiable locations. We cannot, at the present state of the art, computationally analyse more general references to place which potentially have a set of coordinates, such as 'the city' or 'the village', nor can the geoparser generate PNCs for more general geographical areas, such as 'the south'. Given a small text, the former can be added to the analysis manually, using a combination of geoparsing and close reading. However, at the scale of large corpora, this is not feasible. Furthermore, there is currently no systematic way to geoparse references to place which are relative, such as deictic terms like 'here' or 'over there'. Future work to test our findings could, in the first instance, involve performing GTA on additional search terms associated with poverty and could also involve more close reading of texts to uncover deictic references.

Acknowledgements

This work was partly funded by the ESRC Centre for Corpus Approaches to Social Science (CASS): Grant number ES/K002155/1. We would also like to acknowledge the contributions of Andrew Hardie, deputy director for CASS, who has maintained the geoparser and helped to facilitate our analysis of *The Times* corpora.

Notes

- 1 The asterisk in the query term denotes 'zero or more characters', so this particular query will return instances of *poverty*, *poverty-laden*, etc.
- 2 As a result of poor OCR and other corpus formatting issues, we were not able to include texts from a small subset of data for each of the following years: 1971, 1983, 1986, 1988, 2000, 2001, 2005–2007.

Further reading

- 1 Atkinson, P. and Gregory, I. (2017). Finding child welfare in Victorian newspapers: An exercise in corpus-based discourse analysis. *Journal of Interdisciplinary History* 48: 159–186.
This article presents a critical reflection on the use of a large digital archive of Victorian newspapers to explore the discourse around a topic in nineteenth-century history, namely public health.
- 2 Donaldson, C., Gregory, I.N. and Taylor, J.E. (2017). Locating the beautiful, picturesque, sublime and majestic: Spatially analysing the application of aesthetic terminology in descriptions of the English Lake District. *Journal of Historical Geography* 56: 43–60.
This article uses geographical text analysis to explore the way that landscape aesthetics were represented prior to 1900 using a corpus of writing about the English Lake District
- 3 Hitchcock, T. (2013). Confronting the digital: Or how academic history writing lost the plot. *Cultural and Social History* 10: 9–23.
A highly opinionated article that critiques the ways that historians are making use of the digitised sources that are available to them.
- 4 Paterson, L.L. and Gregory, I. (2019). *Representations of poverty and place: Using geographical text analysis to understand discourse*. Basingstoke: Palgrave.
A detailed guide to the use of geographical text analysis using modern newspapers to study poverty.

- 5 Pumfrey, S., Rayson, P. and Mariani, J. (2012). Experiments in 17th century English: Manual versus automatic conceptual history. *Literary and Linguistic Computing* 27(4): 395–408.
This is a pioneering text on the use of corpus linguistics in historical research. It focuses on the changing use of the language of science and experimentation in the early modern period.

References

- Adolphs, S. (2006). *Introducing electronic text analysis*. London: Routledge.
- Atkinson, P. and Gregory, I. (2017). Finding child welfare in Victorian newspapers: An exercise in corpus-based discourse analysis. *Journal of Interdisciplinary History* 48: 159–186.
- Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.
- Baker, P., Gabrielatos, C. and McEnery, T. (2014). Sketching Muslims: A corpus driven analysis of representations around the word ‘Muslim’ in the British press 1998–2009. *Applied Linguistics* 34: 255–278.
- Bauman, Z. (2005). *Work, consumerism, and the new poor*, 2nd ed. Maidenhead: Open University Press.
- Boonstra, O., Breure, L. and Doorn, P. (2004). *Past, present and future of historical information science*. The Hague: DANS.
- Brake, L. and Demoor, M. (2009). *Dictionary of nineteenth-century journalism*. London: British Library Publishing Division.
- Chrisman, N. (2002). *Exploring geographic information systems*, 2nd ed. New York: John Wiley.
- Diggle, P., Pingping, Z. and Durr, P. (2005). Nonparametric estimation of spatial segregation in a multivariate point process: Bovine tuberculosis in Cornwall, UK. *Journal of the Royal Statistical Society Series A* 54: 645–658.
- Donaldson, C., Gregory, I.N. and Taylor, J.E. (2017). Locating the beautiful, picturesque, sublime and majestic: Spatially analysing the application of aesthetic terminology in descriptions of the English lake district. *Journal of Historical Geography* 56: 43–60.
- Dorling, D. (1997). *Death in Britain: How local mortality rates have changed: 1950s-1990s*. York: Joseph Rowntree Foundation.
- Dorling, D. (2013). *The 32 stops – lives on London’s central line*. London: Penguin.
- Dorling, D., Mitchell, R., Shaw, M., Orford, S. and Davey Smith, G. (2000). The ghost of Christmas past: Health effects of poverty in London in 1896 and 1991. *British Medical Journal* 321: 1547–1551.
- Firth, J.R. (1957). *A synopsis of linguistic theory*. Oxford: Blackwell.
- Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2000). *Quantitative geography: Perspectives on spatial analysis*. London: Sage.
- Gabrielatos, C. and Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996–2005. *Journal of English Linguistics* 36: 5–38.
- Gregory, I. (2009). Comparisons between geographies of mortality and deprivation from the 1900s and 2001: Spatial analysis of census and mortality statistics. *British Medical Journal* 339: 676–679.
- Gregory, I. and Donaldson, C. (2016). Geographical text analysis: Digital cartographies of lake district literature. In D. Cooper, C.E. Donaldson and P. Murrieta-Flores (eds.), *Literary mapping in the digital age*. Abingdon: Routledge, pp. 67–87.
- Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S. and Ball, J. (2010). Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A* 368: 3875–3889.
- Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17: 380–409.
- Hardie, A. and McEnery, T. (2009). Corpus linguistics and historical contexts: Text reuse and the expression of bias in early modern English journalism. In R. Bowen, M. Möbärg and S. Ohlander (eds.), *Corpora and discourse – and stuff: Papers in honour of Karin Aijmer* Gothenburg Studies in English 96. Göteborg: Acta Universitatis Gothoburgensis, pp. 59–92.

- Heywood, I., Cornelius, S. and Carver, S. (2006). *An introduction to geographical information systems*, 3rd ed. Harlow: Pearson.
- Hitchcock, T. (2013). Confronting the digital: Or how academic history writing lost the plot. *Cultural and Social History* 10: 9–23.
- Holley, R. (2009). How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine* 15(3–4). Available at: www.dlib.org/dlib/march09/holley/03holley.html.
- Jockers, M.L. (2013). *Macroanalysis: Digital methods and literary history*. Urbana: University of Illinois Press.
- Johnson, S., Culpeper, J. and Suhr, S. (2003). From ‘politically correct councillors’ to ‘Blairite nonsense’: Discourses of political correctness in three British newspapers. *Discourse and Society* 14: 28–47.
- Joulain-Jay, A.T. (2017). *Corpus linguistics for history: The methodology of investigating place-name discourses in digitised nineteenth-century newspapers*. Unpublished PhD thesis, Lancaster University. Available at: <http://eprints.lancs.ac.uk/88671/>.
- Katz, M.B. (2013). *The undeserving poor*, 2nd ed. Oxford: Oxford University Press.
- Kirschenbaum, M.G. (2010). What is digital humanities and what’s it doing in English departments? *ADE Bulletin* 150: 55–61.
- Lloyd, C.D. (2007). *Local models for spatial analysis*. Boca Raton: CRC Press.
- Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (2015). *Geographic information science and systems*, 4th ed. London: Wiley.
- Mack, J. and Lansley, S. (1985). *Poor Britain*. London: George Allen and Unwin.
- Moretti, F. (2013). *Distant Reading*. London: Verso.
- McEnery and Hardie 2011 cited in text but does not appear in reference list.
- McEnery, T. and Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
- Murrieta-Flores, P., Baron, A., Gregory, I., Hardie, A. and Rayson, P. (2015). Automatically analysing large texts in a GIS environment: The registrar general’s reports and cholera in the nineteenth century. *Transactions in GIS* 19: 296–320.
- Mussell, J. (2012). *The nineteenth-century press in the digital age*. Basingstoke: Palgrave Macmillan.
- Nicholson, B. (2010). The digital turn: Exploring the methodological possibilities of digital newspaper archives. *Media History* 19: 59–73.
- Nicholson, B. (2012). Counting culture; or, How to read Victorian newspapers from a distance. *Journal of Victorian Culture* 17: 238–246.
- Paterson, L.L. and Gregory, I. (2019). *Representations of poverty and place: Using geographical text analysis to understand discourse*. Basingstoke: Palgrave Macmillan.
- Piachaud, D. and Webb, J. (2004). Changes in poverty. In H. Glennerster, J. Hills, D. Piachaud and J. Webb (eds.), *One hundred years of poverty and policy*. York: Joseph Rowntree Foundation, pp. 29–27.
- Porter, C., Atkinson, P. and Gregory, I. (2015). Geographical text analysis: A new key to nineteenth-century mortality. *Health and Place* 36: 25–34.
- Pumfrey, S., Rayson, P. and Mariani, J. (2012). Experiments in 17th century English: Manual versus automatic conceptual history. *Literary and Linguistic Computing* 27(4): 395–408.
- Putman, L. (2016). The transnational and the text-searchable: Digitized sources and the shadows they cast. *American Historical Review* 121: 377–402.
- Rowlingson, B. (2015). *Spatial segregation analysis*. Personal Communication.
- Rowntree, B.S. and Lavers, G.R. (1951). *Poverty and the welfare state: A third social survey of York dealing only with economic questions*. London: Longman.
- Rupp, C.J., Rayson, P., Gregory, I., Hardie, A., Joulain, A. and Hartmann, D. (2014). Dealing with heterogeneous big data when geoparsing historical corpora. In *Proceedings of the 2014 IEEE conference on big data*, Washington, DC, pp. 80–83.
- Schreibman, S., Siemens, R. and Unsworth, J. (2014). *A companion to digital humanities*. Oxford: Blackwell.

- Tanner, S., Munoz, T. and Ros, P.H. (2009). Measuring mass text digitization quality and usefulness: Lessons learned from assessing the OCR accuracy of the British library's 19th century online newspaper archive. *D-Lib Magazine* 15(7–8). Available at: www.dlib.org/dlib/july09/munoz/07munoz.html.
- Terras, M. (2016). A decade in digital humanities. *Journal of Siberian Federal University: Humanities & Social Sciences* 7: 1637–1650.
- Thomas, B. and Dorling, D. (2011). *Bankrupt Britain: An atlas of social change*. Bristol: Policy Press.
- Townsend, P. (1979). *Poverty in the United Kingdom: A survey of household resources and standards of living*. Harmondsworth: Penguin.

English language and philosophy

Jonathan Tallant and James Andow

Introduction

There are many different views about the nature of philosophical enquiry. On many of these views, there is a significant philosophical contribution to be made by empirical enquiry into the ways that language is used in the real world by what we might call ‘ordinary language users’. However, philosophers have rarely taken up the opportunity to study language systematically and empirically. The digital resources now available for studying language use in the corpus and other approaches in the digital humanities afford a real opportunity for philosophers to do just this.

This chapter addresses a number of questions. Why would evidence about language use be helpful for philosophers? What resistance might there be to this idea? How might this resistance be reflected in the ways that philosophers make use of corpus research and other approaches in the digital humanities? We begin with an overview of some of the ways in which philosophical enquiry has been carried out.

Background

In this section we outline two popular conceptions of philosophical enquiry and argue that those who conceive of philosophy in any of these ways ought to accept that systematic empirical enquiry into language use has an important contribution to make in philosophy.

Philosophers investigate a range of topics: morality, freedom, identity, existence, time, knowledge, gender and so on. When philosophers think about things like (for instance) freedom, and when they try to provide answers to questions like ‘What is it to be free?’ it seems important that what they mean by ‘free’ has got some connection with ordinary thought and talk about freedom.¹ If the philosopher ends up talking about some completely new concept that isn’t that of freedom, but is somewhat close to it (unimaginatively, let us call it ‘schmeedom’), then there is a sense that the philosopher will have missed the point. We started off asking about freedom, and now the philosophers are off talking about something else – schmeedom (whatever that is and however, if at all, it is connected to the concept of freedom). If we wanted to know about what it is to be free, how that connects with morality, politics,

science and so on, then talking about schmeedom just isn't going to help. With this background in mind, let's now explore some of the ways in which philosophical enquiry has proceeded.

Conceptual analysis

Conceptual analysis is an attempt to analyse concepts. Concepts are, in the relevant sense, constituents of thought. To analyse them is to break them down, to understand their structure in more basic terms. The idea that conceptual analysis is central to philosophical enquiry can be traced back to Socrates (Beaney 2015).

One oft-discussed example is the (flawed) analysis of knowledge as justified true belief. The idea was that the concept of knowledge might be analysed in terms of the more basic concepts of justification, belief and truth. Were the analysis correct, then we could correctly say that a subject, S, knows that p, if and only if, S has a justified true belief that p.

One traditional assumption was that a philosophical analysis of a concept would produce something like a definition, providing a set of necessary and sufficient conditions. Providing something like a definition is, however, just one example of how one might attempt to analyse a concept in terms of something more basic. Following psychologically informed criticism, philosophers have become more open to the idea that the result of analysing a concept might be something other than a definition, for example, the concept might take the form of a prototype, exemplar or theory (see Laurence and Margolis 1999).² It is important to register that even among proponents of conceptual analysis, there are different views about where the project of conceptual analysis should get you (Goldman 2007). Some take the endpoint to be an articulation of the concept of one's linguistic community, whereas others take the endpoint to concern a universally shared concept, or to reveal the nature of the relevant phenomenon itself (e.g. knowledge, via revealing the concept of knowledge).

Now we said earlier that there was a close tie between language use and philosophical methodology. So, what's the connection between analysing concepts and language use? Philosophers will have various opinions about this. We're going to play safe and simply set aside two extreme, implausible positions.

- (1) Language use and concepts have nothing to do with each other.

It might seem simply obvious that this is an implausible position. However, it is worth articulating why.

Language expresses thoughts, and concepts are constituents of thoughts. It would therefore be very strange if language in no way reflected the nature of our concepts. For instance, if a philosopher's analysis of a concept bore no resemblance to the way anyone ever used the word 'knowledge' then it wouldn't be an analysis of the concept of *knowledge*. The analysis would have failed.

If that's right, then there should be scope for systematic empirical investigation into real language use to provide evidence which can inform the project of analysis.

This said, it is also important for us to note that (2) is also completely implausible.³

- (2) A community's concept of knowledge can be simply 'read off' their use of certain words.

To illustrate the problem with this approach, let us use an example. This understanding of the relation between words and language is committed to the idea that when a community

apply the word ‘fish’ to something this is to be interpreted as a correct application of their concept. But that’s implausible, and wouldn’t permit a community to *discover* that, after all, they’d been wrong to think the concept of fish extended to whales. We cannot simply ‘read off’ which applications of a concept are correct from the use of language within a community. It is perfectly possible for communities to just get it wrong. Whatever the relation between language production and conceptual content, it is not as straightforward as (2) would suggest.

Our claim, then, is that there is scope for corpus research, and other lines of research within the digital humanities, to discover previously unknown facts about the factors and associations that, for example, underpin people’s use of (for instance) ‘knows’. For example, were a close analysis of a corpus to reveal the widespread phenomena that competent language users were prepared to assign knowledge to an agent, even in a variety of contexts where that agent has no justification for their belief, then that would have served to put pressure on the claim that we can analyse the concept of knowledge *as* the concept of justified true belief.

In pursuing this line, we advocate a position that we think is taken by Rosenbloom (2012). We are not opposed to philosophical methodology as it has been practised. We recognise that it has value and that it helps us to gather insight. Nonetheless, one way to consider the scope of the digital humanities *in general* is as of giving us additional methodological tools with which to approach old questions (as well, of course, as providing us with more questions to ask – more of which at the end of the chapter). Our general approach to philosophy, as incorporated within the digital humanities is largely summarised thus:

The scientific method provides a validated approach for developing insights, but it is not the only method, nor necessarily always the most appropriate method. For any particular domain, and any particular problem within that domain, there may be zero, one, or more methods applicable to them. If we define the strength of a method for a domain or problem as the degree of veracity it can guarantee for the results it generates, good science should in general be pursued via one from among the strongest applicable methods. Using weaker methods when stronger ones are available can be one of the hallmarks of bad science. We need to be careful, though, about what is meant by two methods being applicable to the same problem, and thus the circumstances under which a stronger method should necessarily dominate. If two methods can yield the same insights and one provides more assurance with respect to these insights, then the two are applicable to the same problem and the stronger should be preferred. However, if the problems are nominally the same but the two methods provide different insights about it, then the weaker method may still be of value, and the problems they tackle are in an important sense different.

The potential diversity of appropriate methods, both within and across domains, does suggest a form of methodological pluralism in which multiple methods may be necessary to increase our understanding of individual domains (Rosenbloom 2012: 7–8).

Not every philosophical question is well treated by the more scientifically oriented aspects of the digital humanities. That’s fine. Nonetheless, the methods that we can draw upon – as we shall show in a moment – do allow us to generate additional insight and give us more data with which to approach long-standing problems. Our primary aim in this chapter is to demonstrate that this is so and to make the case for the inclusion of methodologies from the digital humanities within philosophical enquiry.

Intuition-based enquiry

Many contemporary conceptions of philosophical enquiry give an important role to something they call ‘intuitions’. Philosophers variously treat intuitions as data, as infallible

evidence, as defeasible or *pro tanto* evidence, as inspiration and so on. We shall group all these under one heading. What justifies treating all these approaches under one banner is that they all in some sense treat philosophical enquiry to be limited or guided by what is congruent with our intuitive judgements about philosophically interesting things – knowledge, morality, freedom and so on.⁴ The motivation for this way of approaching philosophy can be somewhat similar: if a philosopher's account of justice were to be completely divorced from our intuitive ways of thinking about justice, it would be unclear what made the account an account of *justice* rather than something else.

What is an intuition? Here, again, we encounter a variety of opinion. We can't hope to do justice to all of them.⁵ We think we can say, however, that most have in mind something like a fairly quick, non-inferential, pre-theoretic judgement about whether something applies. Sometimes a reliance upon intuitions is cast explicitly in linguistic terms – 'intuitively, we wouldn't say that such a person was free' – but often it is not. A fairly typical idea might be that if a theoretical position entails a picture of the world which runs completely counter to our intuitive way of thinking, then this counts against the theory in some way. Given the audience of this chapter, we should probably mention Chomsky. Some philosophers have drawn connections between philosopher's use of 'intuition' and that in linguistics post-Chomsky (Hintikka 1999). However, Chomsky's philosophical inspirations notwithstanding, in the majority of cases, contemporary philosophers' use of 'intuition' does not indicate anything like judgements of grammaticality, nor reflect a belief in a special faculty of intuition analogous to a universal grammar. In many cases, philosophers' use of 'intuition' doesn't indicate any strong commitment to a distinctly *a priori* mode of enquiry.

There's no straightforward consensus concerning the connection between intuition use and language use. However, we're inclined to endorse a similar line of reasoning to that considered earlier when talking about conceptual analysis: the idea that what people say can provide no insight into their intuitive ways of thinking about the world is pretty farfetched. If ordinary intuitive ways of looking at the world are thought to place limitations on which philosophical theories are plausible, or provide evidential support for or against theories, then it looks like corpus linguistics can have a hand to play. For, of course, what corpus linguistics can do for us is to reveal precisely the contexts in which people assert various different claims and so, on the assumption that people do not typically make widespread use of turns of phrase that are 'counterintuitive' corpus linguistics can give us some insight into their intuitions about whether particular uses of a term are 'intuitive' or not.

Current contributions and research

To remind the reader: our aim here is to make the case that philosophy has something to gain from broadening its scope and including particular methodologies from the wider digital humanities. To show that this is the case, we want to consider two particular cases. Methodologically we are looking to show particular cases where digital methods have provided value and then look to generalise from these cases.

The two case studies we consider here concern papers which we wrote (although not together). Of course, we are neither the first to suggest that the philosophical enquiry might benefit from looking at language use in a systematic way nor the first to put such recommendations into practice (the Oslo school did both – see e.g. Naess 1938). However, these two cases studies help make clear a number of points that we want to bring out and that demonstrate the further insight that the digital humanities can bring to philosophical questions. The first is to help make clear the contribution that both quantitative and qualitative approaches

to looking at usage can make to philosophical debate. The second is to demonstrate that examination of language usage, in addition to helping those engaged in conceptual analysis or intuition-using philosophy, can help advance our understanding of philosophical methods themselves – this metaphilosophical theme is one we'll pick up again later.

Exploding intuitions

Andow (2015b) examines an aspect of current debate about the methods philosophers use. A number of high-profile commentaries on the philosophical methods have noted that philosophers use the word 'intuition' a lot more than they used to. Andow presents a quantitative study to investigate this phenomenon – to see what might underpin this change, and indeed to confirm whether there has been a change.

Andow uses JSTOR as a corpus of academic writing. Unlike other available bodies of text, JSTOR allows for fine-grained diachronic study, as articles are tagged by publication year. JSTOR also allows for a comparison of academic disciplines, as articles are tagged by subject areas. Andow was able to track the proportion of articles in each discipline by decade which made use of one or more of the following terms 'intuit', 'intuition', 'intuitive', 'intuitions', 'counter-intuitive', 'intuitively' and 'counter-intuitively'. The headline finding was that, surprisingly, while the proportion of philosophy articles mentioning at least one of these words has risen dramatically, the same has happened across a broad swathe of academia.

The import of this finding is that whatever does underpin the explosion of 'intuition' in philosophy it is likely not to be something philosophy specific. Indeed, Andow also reports data from the *Time Magazine* corpus and Google Ngrams (fiction) which suggest that the frequency of 'intuition' has risen more generally. This means that any metaphilosophical conclusions one might have been tempted to reach on the basis of noting the increase in use of 'intuition', for example, that there has been an increase in the use of intuition, seem not to be warranted.

Intuitions in physics

This second case study also concerns intuitions and philosophical methodology. Tallant uses the results of a qualitative analysis of the use of the word 'intuition' in physics (Tallant 2013) to make an important point about methods in an area of philosophy called metaphysics (Tallant 2015).⁶

Tallant aims to offer some resistance to those who claim metaphysics is not in good epistemic standing on the basis that metaphysicians use intuitions. Tallant is interested in giving a kind of parity defence: physics uses intuitions, too, and there is no doubting the success of physics, so we shouldn't think metaphysics so badly off *just* because it uses intuitions. Tallant's opponents don't accept this parity. Ladyman and Ross (2007), for instance, claim there are important differences between what is going on when a metaphysician uses the word 'intuition' and when a scientist uses the word.

It is frequently said of, for example, a good physicist that he or she has sound physical intuition. . . . However, the meaning of 'intuitions' in these uses differs sharply from the metaphysician's. The physicist . . . [refers] to the experienced practitioner's trained ability to see at a glance how their abstract theoretical structure probably – in advance of essential careful checking – maps onto a problem space.

(p. 15)

Tallant notes that this move requires an empirical claim for which Ladyman and Ross present no evidence. Tallant uses his own recent empirical work (2013) to challenge Ladyman and Ross's assumptions about use of 'intuition' in physics.

Tallant (2013) makes use of a number of techniques. First, like Andow (2015b), earlier, Tallant presents the basic proportions of articles in a number of top physics journals (*Physical Review Letters*, *Review of Modern Physics*, *Physical Review AE*, *Physical Review Special Topics Energy Beams*) between 2001 and 2010 which use 'intuitive' (including 'intuition' and 'counter-intuitive'). It is around 9 percent. However, Tallant goes beyond these crude numbers and provides some qualitative assessment. Second, Tallant presents a basic thematic analysis of the uses of 'intuitive' in this corpus based on a small subset of articles. Third, Tallant also presents findings relating to one journal in particular (*The Review of Modern Physics*) relating the number of uses of 'intuitive' which fall under each theme. Building on this, Tallant (2015) has analysed the data concerning the use of intuitions in physics and the use of intuitions in metaphysics and argued that it is not at all clear that the arguments against metaphysics go through.⁷

Future directions

Whether one thinks looking at the corpus is a worthwhile activity for a philosopher depends on how one thinks of philosophical enquiry. Earlier, we noted some conceptions of philosophical enquiry which could be quite friendly to this sort of research. However, it is worth noting that these conceptions are themselves not uncontroversial.

One helpful way to make this point is to consider the case of experimental philosophy. Experimental philosophy is a relative newcomer on the philosophical scene. It is around 20 years old at the time of publication. The main characterising feature of experimental philosophy is the idea that systematic empirical investigation into how people think about philosophically interesting things, like freedom, morality and the rest, can help advance philosophical debate about those things (Knobe and Nichols 2008). To that end, experimental philosophers advocate the use of experimental methods from social psychology, and the cognitive sciences more broadly.⁸ Experimental philosophers motivate their use of these methods by appealing to precisely the same kind of considerations that we have given to motivate systematic empirical study of language use. Experimental philosophers too have appealed to popular conceptions of philosophical enquiry such as conceptual analysis and intuition-using philosophy (See Knobe and Nichols 2008).⁹

The lesson we can learn from this concerns the resistance experimental philosophers have faced from some other philosophers (e.g. Cappelen 2012, 2014; Deutsch 2010; Kauppinen 2007). One of the responses that experimental philosophers face is that their work is fundamentally misconceived – that the kind of data they were collecting, for example, what ordinary people think about morality, is irrelevant to what the correct moral theory might be. This response relies upon the idea that conceptions of philosophical enquiry as being conceptual analysis or using intuition are mistaken – or at least that the ways of understanding these projects such that they are amenable to empirical enquiry of the kind advocated by experimental philosophers are mistaken.

Accordingly, advocates of corpus-based methods in philosophy should be prepared for similar resistance. We'll return to this thought shortly. First, however, it is interesting to consider a respect in which this resistance shaped the trajectory of experimental philosophy. This relates to the point we made earlier about the contribution made by our two case studies being predominantly metaphilosophical.

A dominant trend in early experimental philosophy was it was not involved in ‘first-order’ philosophy, for example, using experiments to inform philosophical theorising about morality, but rather in metaphilosophy. A number of much-discussed papers were published that tried to challenge common metaphilosophical assumptions: for example, to make the case that philosophers couldn’t safely assume that their intuitions were reliable or that their intuitions were representative of the population at large.¹⁰ This early metaphilosophical flavour to experimental philosophy should, in some respects, not be too surprising given the nature of the resistance experimental philosophy faced. Metaphilosophical work faced less resistance at least in the sense that it was far less easy for philosophers to dismiss the work as *irrelevant* – the conclusions might be *wrong* but not *irrelevant* – these metaphilosophical experimental studies raised direct challenges to standard philosophical practice and, even if philosophers thought the challenges were mistaken, it was recognised that these studies needed to be discussed. It is thus then not so surprising that there is a metaphilosophical flavour to our two case studies.¹¹

Andow (2015b) and Tallant (2013, 2015) also share another feature in common with early work in experimental philosophy.¹² Another type of study which was prominent in the early days of experimental philosophy was that which picks up on empirical claims made in the philosophical literature. Sometimes these were explicit claims which clearly played the role of premises in an argument (e.g. Nichols et al. 2003). This is what Tallant (2015) did – picking up on Ladyman and Ross’s claim concerning the role of intuition in science. Other times, the relevant empirical claims were not necessarily used as premises in any explicit argument, but nonetheless betrayed a particular underlying understanding of some particular phenomenon (see, e.g. Stich and Weinberg 2001; Knobe 2007’s description of the inspiration for the now-rich literature on whether non-philosophers are compatibilists). This is what Andow did – challenging the idea that the major factors underlying philosophers’ use of the word ‘intuition’ are philosophy-specific factors. These sorts of cases encounter less resistance because the relevance of the empirical work to the philosophical debate is simply obvious: when philosophers make empirical assumptions, these are open to empirical challenge.

The lessons to take from all this is that we should not expect philosophers to start pursuing conceptual analysis using serious corpus research overnight. We should expect similar resistance to the employment of methods from corpus linguistics. But the parallel with experimental philosophy also gives a source of hope. Experimental philosophy has flourished. Further, it has not confined itself to metaphilosophical contributions and testing philosophers’ empirical claims. The bulk of current work in experimental philosophy is engaged in a serious attempt to understand how people *think* about philosophically interesting things. So, we think there is reason to think that – given the right impetus and support – the use of systematic empirical study of actual language use will also be able to flourish and develop serious empirical programmes investigating how people *talk* about philosophically interesting things.¹³

Critical issues

It is well and good for us to argue that there is scope for these tools to be used. There will be issues faced along the way. We wish to discuss some of these here in the interests of trying to argue further in favour of their uptake.

The proponent of use of methods from corpus linguistics (and the digital humanities more generally) in philosophy should be encouraged by the fate of experimental philosophy.

We must also keep in mind certain facts from the recent history of philosophy. One school of thought in philosophy, prevalent in the early-to-mid twentieth century, focused almost exclusively on language use. This school, called ordinary language philosophy, treated the sole task of philosophy as reflecting on linguistic practice. In part for reasons we give earlier, and in part for other reasons, this approach to philosophical enquiry has become markedly less popular in recent years. Philosophers tend to see the words that we use as a guide to how we think reality is (and given the fact that these thoughts appear in successful scientific theories about the world, how the world itself is), but their focus extends beyond the words and into the world itself. We would be sensitive, then, to any criticism that what we are extolling here is an approach to philosophy that *just* focuses on the way that people use words. We take the view that the approach that we are arguing in favour of here is one that incorporates the linguistic data available via (e.g.) corpus linguistics and applies them in a range of different cases. To illustrate, we provide three cases that we think are ripe for exploration via corpus linguistics or other methods that might all fall under the remit of the digital humanities.

The first case is drawn from epistemology. Within epistemology, there are long-standing problems. In particular, there is a question about whether or not knowledge is possible. It seems obvious that it is, of course. For example, you know that you're reading this chapter. But things are not so straightforward. For it seems at least *possible* that you're wrong about that. Isn't it *possible* (in the most expansive sense of the word 'possible') that you're being deceived? *Might* you not be installed in some vat – à la *The Matrix*? It seems you can't rule out the possibility and if you can't rule out the possibility of being a brain in a vat, then it seems false to say that you know that you are reading this chapter.

One popular recent response has been to argue that this is an example of the standards for the term 'knows' applying varying from context to context (for an overview see Rysiew 2011). In normal contexts, it seems obviously correct to say that you know that you are reading the paper. But once we introduce some sceptical possibilities, we 'raise the standards' required for knowledge. The idea, very roughly, is that the term 'knows' functions similarly to the term 'tall': the term tall can be truthfully applied to Jack in some contexts (when Jack is visiting his daughter's primary school class), but it can also be truthfully said of Jack in other contexts that he is 'not tall' (for instance, if he is visiting an NBA team).

We do not propose to delve into the debates about the term 'knows' or to involve ourselves in the intricacies of the different positions available, but we do note that all parties in this debate appeal to linguistic data. It is common enough to see philosophers in this literature claim that within a particularly stylised context, the correct use of a particular term is *x*, where *x* then supports their theory or undermines the theory of their opponent.¹⁴ Where we think that corpus linguistics may get some traction, then, is in determining whether there are such cases 'in the wild' and whether the terms being discussed *are* used as the philosophers claim. We might, thus, move the literature beyond philosophers swapping their linguistic intuitions, and deploy large datasets to determine which usages really are viewed as apt.¹⁵

Our second case is drawn from metaphysics – in particular, the metaphysics of time. One semi-popular view in this literature is presentism – the view that only present objects exist. This view, presentism, is often motivated by the claim that it is particularly intuitive; that it is the view of the 'person on the street' and is particularly commonsensical. To the best of our knowledge, these claims have never been verified. One way to test these claims would be via interviews or surveys – the latter are used regularly enough in the experimental philosophy that we described earlier; interestingly we know of no

one using interviews. Another way in which to test these claims would be via corpus study. For, presumably, if presentism really is the most intuitive view of time, and really is the common-sense view of the ‘person on the street’, then we should expect to see that reflected in their use of language. Very crudely: we might expect to see past objects described as ‘non-existent’; we might expect to see future objects described in such a way as to imply their non-existence. And were we to not find that, it might well put some pressure on the claim that presentism is especially intuitive, or really is the common-sense view of the person on the street.

Our third and final case is drawn from the branch of philosophy that explores propositions. Very roughly a proposition is thought to be that which is expressed by sentences – the meaning of what is said and thought, if you will. (Thus, two sentences in different languages can be used to express the one and the same proposition; the two sentences can have the same meaning.) One debate about the nature of propositions concerns whether different tokens of the same type of sentence, uttered at different times, express different propositions, or whether they express the same proposition – though a proposition that can change with respect to whether it is true or false. Consider the sentence:

(1) Nixon is president

On a temporalist reading, what is expressed by (1) may change its truth-value over time: once it was true, now it is not. Nonetheless, there is but one proposition expressed by (1) at the two times under consideration, and it changes its truth-value. In contrast, on an eternalist reading, the proposition expressed by (1) does not change over time. Rather, at different times the same sentence can be used to express different propositions. Very crudely, we might think that something like the following is going on if eternalism is true. Uttered in 1970, (1) expressed a proposition that had the form: at 1970, Nixon is president. Uttered in 2016, (1) expresses the proposition that has the form: at 2016, Nixon is president. It is clear enough that the latter is false and that the former is true. However, such propositions are, if true at all, eternally true. The proposition at 1970, Nixon is president is true now, it was true in 1970 and it will be true in a thousand years’ time.

One recent piece of work on this topic, due to Mark Richard (2015) explores this issue by considering a range of possible assertions, using these as evidence to favour one of these two views. Throughout the book, where he deals with a range of different issues, Richard draws on the ways that, he thinks, people will think about and respond to the cases that he considers (2015: 116 is especially vivid). In his own terms, here is Richard describing his methodology: 2015: 3 – our italics):

the argument is resolutely empirical: *if we look at the way that we talk about belief, assertion, and so on, we find it poorly understood in terms of transiently true propositions.*

Now, be that as it may, Richard does not engage in any of the tools that the digital humanities might provide us with: he does not deploy corpus linguistics, he does not use surveys, he does not interview respondents. In other words, there are a number of tools that would help us to *look* at the way that we talk about belief, assertion and so on, and these are not being exploited. He is not alone in this field. We think that Richard’s work is very powerful, as is the work of others in this field, but it may be that augmenting it with the resources from the digital humanities would render it even more so (see Tallant forthcoming for further discussion).

Mastering corpus research methods for use in philosophy

Here again we can learn lessons from the fate of experimental philosophy. The general narrative told about experimental philosophy is that it has grown to be very empirically sophisticated and expert from rather less sophisticated and expert beginnings. There are various reasons for this. Early experimental philosophers didn't have an extensive grounding in statistics or survey design. They had to learn on the job. They were creating research traditions and had to develop new experimental paradigms. They had to develop new tools to look at *philosophically* interesting issues which the psychological literature didn't address. Similar issues will face proponents of empirical investigation of language-use in philosophy. For example, we definitely recognise (1) that our corpus case studies discussed earlier have their limitations and (2) that the fact we were newcomers to corpus-based work and we were asking new types of questions held us back in certain regards.

There was no pre-existing corpus structured in such a way that would allow one to address the questions Andow (2015b) addresses and the skills needed to develop the necessary corpus are far beyond a philosophers' usual skillset. While Andow was able to be creative – utilising JSTOR – this means the study has important limitations, e.g. an inability to look closer at different uses (as one can't look within articles except by manually checking some sample of those to which the researcher has institutional access), and a lack of frequency data.¹⁶

The study carried out by Tallant also had certain limitations. As noted earlier, the approach that Tallant took was to search for instances of the words 'intuition' and 'intuitive' in a range of scientific literature. But this approach is, we suggest, rather simplistic. For instance, the approach that Tallant (2013) takes ignores the issue of whether or not these instances of terms are in fact referring to a use of intuition or whether they appear as (for instance) something like a verbal tic (see Cappelen 2012). Thus, on the basis of this evidence Tallant (2013) fails to establish (though admittedly does not take to have established) that intuitions really are being *used* in physics (though Tallant 2015 is more definite in this regard).

Similarly, Tallant fails to track whether or not *other* terms in the scientific literature might stand proxy for 'intuition' in particular contexts. We note, here, that there is discussion in the philosophical literature as to whether or not apparent cognates of 'intuitive' like 'it seems common sense' and 'it seems right to suggest' *really do* all refer to intuition. But as will be familiar to those with a background in corpus linguistics, the norm when engaging in a study such as Tallant's (2013) is to generate a 'coding manual' that allows the researcher to look for a particular term *and* its cognates, thus identifying cases where intuition (or another concept) is being referred to, but in an indirect fashion. All such uses were missed by Tallant (2013). As alluded to earlier, philosophers are interested in whether particular concepts or ideas are being used, and are not interested in cases where words are used with alternative meanings, or cases where, as just suggested, a term might not be referring to a particular concept, but might simply be a product of a particular fashion of writing. It is therefore pressing that philosophers employing the methods from corpus linguistics do so with an appropriate respect for the methodology required by the linguistic approach.

Here another lesson can be learned from experimental philosophy. Experimental philosophy really came into its own through genuine collaboration between philosophers and psychologists, and through the interest and enthusiasm of graduate students whose dissertation projects included empirical elements. For these things, you need the right supportive environment. This includes things like willingness to work collaboratively across disciplinary boundaries, willingness to co-supervise PhD students outside one's own area and an openness and

receptiveness to new methodological approaches. One might then expect the value of corpus linguistics to philosophy to become evident only once the project is properly up and running.

In general, though, these teething issues are to be expected. Though tools and methods from the digital humanities have not fully broken through into philosophy as yet, they have begun to do so in other disciplines. In her discussion of the development of these methods, Terras (2016: 1641) notes that:

the most difficult, intellectual work of applying technology in the humanities still occurs before the chasm has been crossed, in the phase of innovation, and early adoption, where we are looking at the technologies that cross our path and saying ‘how can I use, or develop this tool for use, in my research?’, much like those in the 1950s or 1960s who were coming across university mainframes and asking how best to apply that in the literary and linguistic arena.

And though we take ourselves to be making the case *for* particular uses of these methods within philosophy, it’s clear enough that more work will need to be done. But – and this really is our point here – that work has not yet really been done. In the absence of clear and well-developed discussions in the wider philosophical literature as to how these different methodologies can be deployed, it’s unsurprising that there is not yet a significant body of heavily digitally driven philosophical research.

Worries about chauvinism

The profound questions of philosophy are not supposed to be *English* questions. Historically that should be clear. The big questions of Anglophone philosophy are ones which can trace their discussion through ancient Greek, Latin, Arabic and many of the modern European languages. Moreover, these discussions have analogues in philosophical traditions from around the world, including various African, Indian and Chinese traditions. Questions about existence, human nature and freedom are simply not language-specific questions. So, approaching them solely through usage of English words in English corpora risks adopting an intrinsically chauvinist methodology.¹⁷ Now, of course, this is mitigated a bit by the fact that we would never expect a philosopher to be using corpora as their sole resource. And there are probably some philosophers who would be happy restricting philosophical enquiry to the analysis of culturally specific ways of thinking and talking about such things.¹⁸ Nonetheless, for many philosophers, the worry will be there.

This worry is in some ways analogous to worries that many have raised about the demographic breakdown of participants in psychology experiments (Henrich et al. 2010a, 2010b). Similar worries extend to many studies in experimental philosophy. It is also analogous to certain worries that various people have raised about relying on intuitions in philosophy – where as a contingent matter of fact relying on intuitions typically means relying on the intuitions of a very demographically homogeneous group. Given that the worry is analogous, one might expect the solution to be the same. In these other cases, the solution (or part of it) is typically to diversify your sample in some way.

In experimental philosophy and psychology, the solution is to make sure one has a diverse sample or to encourage cross-cultural studies and replication attempts.¹⁹ In traditional philosophy, one can take extra steps to encourage those from under-represented demographic groups into the profession.²⁰ In the case of corpus research, the analogous solution brings difficulties. How does one diversify the sample? The obvious way is to look at multiple languages. This brings obvious practical difficulties. But, more importantly, the problem of

a non-diverse sample may already be embodied in the existing materials, for example, there are structural problems in how representative corpora are in the first place. For obvious reasons, corpora of English are much better established (even than other European languages, e.g. French) (Sharoff 2006). There might also be worries about the representation of the language-use of various groups within the corpora philosophers might use. Seildhofer (2012) notes the language compiled in BNC/COCA is effectively defined as “what has been produced by so-called educated and predominantly British and American (etc.) native speakers” (p.136). Auer et al. (2014) note “the upper layers of society and men are over-represented in diachronic corpora” (p.10). The underrepresentation of the language-use of women in corpora has been noted from the Early Modern period (Culpeper & Kytö 2010, p.26), to internet corpora (Sharoff 2006, p.444). Of course, there are sophisticated techniques available to help create corpora with better representation of, for example, female writers and to allow for cross-linguistic work. For instance, the use of web crawling (*à la* the WaCky initiative – Baroni et al. 2009) and the creation and use of parallel and comparable corpora (see McEnery and Xiao 2007). However, philosophers probably oughtn’t to be expected to be able to jump in at the deep end. It seems likely that those interested in bringing corpus-based methods to philosophy will have to content themselves with more established, freely available and straightforward resources for the time being. In many cases, this will likely mean focusing on use of English. Nonetheless, making sure that their findings don’t simply reflect quirks of *English* usage will be something such researchers will need to look out for.²¹

And, more generally, we must recognise that philosophy is not unique and will follow a well-trodden path. If we can derive value from the digital humanities, then it seems likely that we will have to continue to do so cautiously in the longer term.

The Gartner Hype Cycle . . . looks at how technologies are launched, mature, and are applied (so people know when to invest). The premise of this is that when technologies are first triggered, everyone thinks they are going to be the Next Big Thing, and so they reach ‘the peak of inflated expectations’, before crashing down into a ‘trough of disillusionment’ when those adopting them realise they are not that great at all. It is hard work to get technologies up the ‘slope of enlightenment’ where useful, useable applications are found.

(Terras 2016: 1642–1643)

As we say, we assume that philosophy will be no different. It will be difficult to generate uptake of new methods and, if we can, this may well lead to an inflation of expectations – a risk that all new technologies carry with them. Our development of these methods will thus need to be rigorous and cautious. Nonetheless, it is clear that philosophy is a discipline that has much to learn and much to benefit from when it come to the digital humanities.

Indeed, we want to move on with a note of optimism. The methodology that we are proposing, here, has promise. While there may be worries about focusing on English and the demographic representativeness of widely available corpora, it is important to note that, while imperfect, use of corpora is almost immediately an improvement on, for example, the standard philosophical practice of interrogating only one’s own and other philosophers’ ideas about ‘what people would say’. Large and widely available corpora such as COCA and BNC incorporate usage from speech, fiction, popular magazines, newspapers, etc., and allow one to focus on particular types of use. This gives *some* traction into the question of how ideas and concepts are used a long way outside the academy. Similarly, one might think that access to corpora of web-based usage (e.g. UKWaC) and particularly social media (e.g. HERMES) would allow another insight into how language is used (in highly non-academic

contexts). These are exciting possibilities. If we can take advantage of them, then this will likely allow philosophers to explore language-use in ways which they have not been able to, to this point.

Summary

The potential for corpus-based research to contribute to philosophical enquiry is clear. Many ways of conceiving of philosophical enquiry make it amenable to empirical enquiry. However, the road to the use of such methods being accepted in philosophy won't be a smooth one. There are important issues to be addressed and metaphilosophical questions to be asked and answered. Just like experimental philosophy, corpus-based research in philosophy should be expected to encounter heated resistance and raise heated discussions about the nature of philosophical enquiry. If the fate of corpus-research in philosophy were to be anything like that of experimental philosophy, corpus research will probably at first contribute largely at the metaphilosophical level (raising challenges and spurring debate) and in putting philosophers' empirical claims to the test before becoming established as a field and developing clear research programmes of its own.

We close on a rather different note. For many subjects, there is a philosophy of that subject: physics, chemistry, biology, psychology, mathematics, economics – the list goes on. What mainstream philosophy has not yet managed to produce is a substantial philosophy of the digital humanities (or of any of the sub-fields contained therein). Of course, that is not to say that these areas are devoid of philosophical interest or that no good philosophy is being done in those fields. We make no such claims. Indeed, tentatively, we think that there are areas where careful philosophical analysis *may* prove useful – though we wish to only suggest such a conclusion. One central question to which philosophy may be able to add something concerns the question of when a corpus is representative, and what it is representative *of*. Indeed, more generally, applied linguistics itself is a field that encompasses a considerable variety of approaches and stepping off points (see, e.g. Scholz et al. 2016). Making appropriate use of philosophical tools (conceptual analysis, for instance) may ultimately prove a useful way of making progress here. There are considerable bodies of work within the philosophical canon that explore what *it is* to represent something; what representation *itself is*. Deploying these notions may well be of use in the development and refinement of new corpus methods.

Notes

- 1 We'll employ the convention of using single quotes only when we want to say something about the words or expressions themselves.
- 2 For a recent, psychologically informed account of how conceptual analysis might work, see Machery (2017).
- 3 Although, naturally, were you to think that, then you would need little persuasion of the main claims in this chapter.
- 4 Some ways of cashing out the project of conceptual analysis also give an important role to intuitions. But we'll treat the two separately here.
- 5 For a brief introduction to some of them, see Jenkins (2014), Pust (2014) and Cappelen (2012).
- 6 Tallant also draws on an ethnographic study by Marton et al. (1992) of interview data from Nobel laureates and survey data from a study by Martinson et al. (2005) concerning scientists' use of 'gut feelings'.
- 7 Again, there are subtleties to this debate, and Tallant (2013, 2015) offers only a partial defence. Please turn to the original papers for the full story. This brief account suffices for our purpose here.

- 8 There is room for debate about how exactly to characterise experimental philosophy. Some have advocated conceptions broad enough such that the approaches we are advocating here would snug-gle in under the umbrella. We are inclined to think there is reason to keep empirical approaches focused on language use separate, however.
- 9 Although, one of us has argued that this is problematic (Andow 2016).
- 10 In fact, more generally, one of the major influences of experimental philosophy has been to lead philosophers to think carefully about their methodological assumptions. For example, various parties have come out in force articulating conceptions of intuition-using philosophy which do not rely on assumptions that philosophers' intuitions would be representative of some wider population (e.g. Williamson 2007). One of the latest trends is to argue that while it might look like philosophers use intuitions as evidence, in fact they don't do anything of the kind (Deutsch 2015; Cappelen 2012).
- 11 This early trend in experimental philosophy has not persisted. This kind of metaphilosophical work has now largely given way to other projects. Only 1.1 percent of experimental philosophy papers published in the period 2009–2013 were of this kind (Knobe 2016).
- 12 This feature is also shared by an example of research in the history of philosophy that uses the tools of corpus research. Edward Slingerland and colleagues use corpus tools to evaluate empirical claims about a historical philosophical corpus (e.g. Slingerland 2013; Slingerland and Chudek 2011; Slingerland et al 2017). They use quantitative tools to reanalyse the corpus of ancient Chinese philosophy to challenge a standard narrative that the idea of mind–body dualism is absent in ancient Chinese philosophy (and perhaps peculiar to Western philosophy).
- 13 One space to watch is the work of Alfano and colleagues. In recent articles, they make the case for greater use of data mining and data visualisation in philosophy (Alfano and Higgins 2019) and put it into practice (Alfano et al. 2018).
- 14 Indeed, Hansen (2014) construes this debate as a contemporary form of ordinary language philosophy.
- 15 This is a particularly promising area because it is somewhat common for epistemologists to recognise the potential import of data from the corpus and to make some limited use of this (e.g. Nagel 2013, n.22), and there is already a big experimental philosophy literature in this area (see Buckwalter 2012 for a survey).
- 16 In fact, Andow (2015b) did suggest several potential explanations of his results, and he did follow up this study with a qualitative study of intuition talk across contemporary academic use (Andow 2015a). However, due to a lack of familiarity with applied linguistics, one potential explanation was overlooked. This, and its metaphilosophical implications, are now explored in Andow (2017).
- 17 Worries in this vicinity seem less urgent and perhaps even irrelevant for corpus research, where the primary objective is to better understand something about language, a specific language or a specific culture.
- 18 Although, even so-called 'ethno-philosophers' tend to place emphasis on using comparing and contrasting different ways of thinking (e.g. Maffie 2002).
- 19 Stephen Stich and Edouard Machery have a project underway to help do just this – *Intellectual Humility and Cultural Diversity in Philosophy: An Examination of the Extent and Implications of Cultural Diversity in Philosophical Intuition* (see thethrivecenter.org/research/research-projects/the-science-of-intellectual-humility).
- 20 Various projects are underway to help do this. For example, see *Minorities and Philosophy* (mapforthe-gap.com) and *Gendered Conference Campaign* (feministphilosophers.wordpress.com/gendered-conference-campaign).
- 21 Note again, that the metaphilosophical bent of Andow and Tallant's case studies means that they are (to some extent) isolated from such worries given the nature of the academic writings relevant to their research questions as their questions do concern specific communities.

Future reading

- 1 Bluhm, R. (2016). Corpus Analysis in Philosophy In Martin Hinton (ed.), *Evidence, Experiment and Argument in Linguistics and the Philosophy of Language*. Peter Lang, pp. 91–109.
- 2 Caton, J.N. (2020). Using linguistic corpora as a philosophical tool. *Metaphilosophy* 51(1): 51–70.
- 3 Scholz, Barbara C., Pelletier, Francis Jeffry and Pullum, Geoffrey K. (2016). *Philosophy of Linguistics, The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.). Available at: <https://plato.stanford.edu/archives/win2016/entries/linguistics/>

References

- Alfano, M. and Higgins, A. (2019). Natural language processing and semantic network visualization for philosophers. In E. Fischer and M. Curtis (eds.), *Methodological advances in experimental philosophy*. London: Bloomsbury.
- Alfano, M., Higgins, A. and Levernier, J. (2018). Identifying virtues and values through obituary data-mining. *The Journal of Value Inquiry* 52(1): 59–79.
- Andow, J. (2015a). How distinction is philosophers' intuition talk? *Metaphilosophy* 46(4–5): 515–538.
- Andow, J. (2015b). How 'intuition' exploded. *Metaphilosophy* 46(2): 189–212.
- Andow, J. (2016). Intuitions (a critical notice of Rowbottom and Booth. eds. *Intuitions*. OUP.). *Analysis* 76(2): 232–246.
- Andow, J. (2017). Intuition-talk: Virtue or vice. *Philosophia* 45(2): 523–531.
- Auer, A., Laitinen, M., Gordon, M. and Fairman, T. (2014). An electronic corpus of letters of artisans and the labouring poor (England, c. 1750–1835): Compilation principles and coding conventions. In V. Lieven, K. Davidse, C. Gentens and D. Kimps (eds.), *Recent advances in corpus linguistics: Developing and exploiting corpora*. Amsterdam and New York: Rodopi, pp. 7–29.
- Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3): 209–226.
- Beaney, M. (2015). Analysis. In E.N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, Spring 2015 ed. Available at: <https://plato.stanford.edu/archives/spr2015/entries/analysis/> (Accessed 9 April 2019).
- Buckwalter, W. (2012). Non-traditional factors in judgments about knowledge. *Philosophy Compass* 7(4): 278–289.
- Cappelen, H. (2012). *Philosophy without intuitions*. Oxford: Oxford University Press.
- Cappelen, H. (2014). X-Phi without intuitions? In A.R. Booth and D.P. Rowbottom (eds.), *Intuitions*. Oxford: Oxford University Press.
- Culpeper, J. and Kytö, M. (2010). *Early modern English dialogues: Spoken interaction as writing*. Cambridge: Cambridge University Press.
- Deutsch, M. (2010). Intuitions, counter-examples, and experimental philosophy. *Review of Philosophy and Psychology* 1(3): 447–460.
- Deutsch, M. (2015). *The myth of the intuitive: Experimental philosophy and philosophical method*. London: MIT Press.
- Goldman, A.I. (2007). Philosophical intuitions: Their target, their source, and their epistemic status. *Grazer Philosophische Studien* 74(1): 1–26.
- Hansen, N. (2014). Contemporary ordinary language philosophy. *Philosophy Compass* 9(8): 556–569.
- Henrich, J., Heine, S.J. and Norenzayan, A. (2010a). Most people are not weird. *Nature* 466(7302): 29.
- Henrich, J., Heine, S.J. and Norenzayan, A. (2010b). The weirdest people in the world? *Behavioral and Brain Sciences* 33(2–3): 61–83.
- Hintikka, J. (1999). The emperor's new intuitions. *The Journal of Philosophy* 96(3): 127–147.
- Jenkins, C. (2014). 'Intuition', intuitions, concepts and the A priori. In A. Booth and D. Rowbottom (eds.), *Intuitions*. Oxford: Oxford University Press.
- Kauppinen, A. (2007). The rise and fall of experimental philosophy. *Philosophical Explorations* 10(2): 95–118.
- Knobe, J. (2007). Experimental philosophy. *Philosophy Compass* 2(1): 81–92.
- Knobe, J. (2016). Experimental philosophy is cognitive science. In J. Sytsma and W. Buckwalter (eds.), *A companion to experimental philosophy*. Oxford: Blackwell.
- Knobe, J. and Nichols, S. (2008). An experimental philosophy manifesto. In J. Knobe and S. Nichols (eds.), *Experimental philosophy*, vol. 2. Oxford: Oxford University Press, pp. 3–14.
- Ladyman, J. and Ross, D. (2007). *Every thing must go*. Oxford: Oxford University Press.
- Laurence, S. and Margolis, E. (1999). Concepts and cognitive science. In S. Laurence and E. Margolis (eds.), *Concepts: Core readings*. London: MIT Press, pp. 3–81.

- Machery, E. (2017). *Philosophy with its proper bounds*. Oxford: Oxford University Press.
- Maffie, J. (2002). Why care about Nezahualcoyotl? Veritism and Nahua philosophy. *Philosophy of the Social Sciences* 32(1): 71–91.
- Martinson, B.C., Anderson, M.S. and De Vries, R. (2005). Scientists behaving badly. *Nature* 435(7043): 737.
- Marton, F., Fensham, P. and Chaiklin, S. (1994). A nobel's eye view of scientific intuition: Discussions with the nobel prize-winners in physics, chemistry and medicine (1970–86). *International Journal of Science Education* 16(4): 457–473.
- McEnery, A. and Xiao, R. (2007). Parallel and comparable corpora: What is happening. Incorporating corpora. *The Linguist and the Translator*: 18–31.
- Naess, A. (1938). 'Truth' as conceived by those who are not professional philosophers. Oslo: I Kom-misjon Hos Jacob Dybwad, pp. 12–15.
- Nagel, J. (2013). Knowledge as a mental state. *Oxford Studies in Epistemology* 4: 275–310.
- Nichols, S., Stich, S. and Weinberg, J.M. (2003). Metaskepticism: Meditations in ethnoepistemology. In S. Luper (ed.), *The skeptics*. Aldershot: Ashgate, pp. 227–247.
- Pust, J. (2014). Intuition. In E.N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, Fall ed. Available at: <https://plato.stanford.edu/archives/fall2014/entries/intuition/> (Accessed 9 April 2019).
- Richard, M. (2015). *Truth and truth bearers: Meaning in context*, vol. 2. Oxford: Oxford University Press.
- Rosenbloom, P. (2012). Towards a conceptual framework for the digital humanities. *Digital Humanities Quarterly* 6(2): 219–233.
- Rysiew, P. (2011). Epistemic contextualism. In E.N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, Winter 2011 ed. Available at: <https://plato.stanford.edu/archives/win2011/entries/contextualism-epistemology/> (Accessed 9 April 2019).
- Scholz, B.C., Pelletier, F.J. and Pullum, G.K. (2016). Philosophy of linguistics. In E.N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, Winter 2016 ed. Available at: <https://plato.stanford.edu/archives/win2016/entries/linguistics/> (Accessed 9 April 2019).
- Seidhofer, B. (2012). Corpora and English as a Lingua Franca. In K. Hyland, C.M. Huat and M. Handford (eds.), *Corpus applications in applied linguistics*. London: Bloomsbury.
- Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni and S. Bernardini (eds.), *Wacky! Working papers on the web as corpus*. Bologna: Gedit, pp. 63–98.
- Slingerland, E. (2013). Body and mind in early China: An integrated humanities – science approach. *Journal of the American Academy of Religion* 81(1): 6–55.
- Slingerland, E. and Chudek, M. (2011). The prevalence of mind – body dualism in early China. *Cognitive Science* 35: 997–1007.
- Slingerland, E., Nichols, N., Neilbo, K. and Logan, C. (2017). The distant reading of religious texts: A 'big data' approach to mind-body concepts in early China. *Journal of the American Academy of Religion* 85(4): 985–1016.
- Stich, S. and Weinberg, J. (2001). Jackson's empirical assumptions. *Philosophy and Phenomenological Research* 62(3): 637–643.
- Tallant, J. (2013). Intuitions in physics. *Synthese* 190: 2959–2980.
- Tallant, J. (2015). Metaphysics, intuitions and physics. *Ratio* 28: 286–301.
- Tallant, J. (Forthcoming). Critical review. *Analysis*.
- Terras, M. (2016). A decade in digital humanities. *Journal of Siberian Federal University: Humanities & Social Sciences* 7: 1637–1650.
- Williamson, T. (2007). *Philosophy of philosophy*. Oxford: Blackwell.

English language and multimodal narrative

Riki Thompson

Introduction

Definitions of the digital humanities (DH) are numerous and expansive, with competing and complementary views on the topic (Borek et al. 2016; Fails and Kelton 2016). Moves to define the field and figure out what ‘counts’ as DH have been going on at conferences and in journals for the last decade. At THATCamp Seattle 2010, the opening session was dedicated to unpacking the various definitions and identifying scholarly areas of interest to manage the logistics of creating sessions around research affinities. As ‘unconference’ participants attempted to define the field, concerns were raised about narrow definitions that potentially excluded scholars whose research brought together the humanities and technology. A year later, Melissa Terras (2011) gave a plenary talk at Interface where she addressed this continuing crisis of inclusion:

It’s something to do with the Humanities, and digital technologies, but what exactly, we are slow – and even reluctant – to pinpoint. The latest definition, ‘Big Tent Digital Humanities’, deliberately obfuscates the focus of the field. . . . Big Tent DH, then, is an ecumenical approach, whilst giving the freedom for individual scholars to explore their own interests, wherever in the research and teaching spectrum they lie.

In the digital age, storytellers have been able to harness technologies to present narratives in new ways. Unlike previous eras that showed a preference for ‘monomodality’ in communication, there is now a movement towards multimodality in which the textual is increasingly integrated with other semiotic modes rather than segregated and used in isolation (Kress and Leeuwen 2001; Machin 2007; Kress 2009). According to Jenkins (2006), we are in an age of media transition and instability, a ‘convergence culture’ in which new and old media interact in a complex relationship. An example of a narrative form that exemplifies the complexity of old and new media is whiteboard animation. This relatively new genre, also known as sketchboard animation, video scribing or fast draw video, is a computer-mediated form in which a sequence of events is animated on a white surface in time with an audio voice-over. With the development of whiteboard animation comes a new mode of narrative

communication in need of examination. This chapter contributes to Big Tent DH scholarship by examining the evolving technology of whiteboard animation as a relatively new digitally mediated storytelling form.

Background

Many of the components of this relatively new narrative form have been around for a very long time. The use of stop-motion photography can be seen in clay-animation films (also known as Claymation) that have been around since 1908, and the process of animating cartoons and combining them with film was pioneered by Disney in the 1930s. Though it is not clear when the first whiteboard animation was developed, the emergence of the genre can be connected to the popularity of the RSAAnimate series that was published online in 2010 when Cognitive Media adapted the public lecture series of Royal Society for the Encouragement of Arts, Manufactures and Commerce (RSA) to ‘help people to discover and learn new stuff through storytelling’ (Cognitive Media 2017).

Some whiteboard animation videos rely on the material practices of drawing and photography, whereas others mimic the material, using digital animation software and video to generate a story. In early whiteboard animation, a cartoonist drew the story frame by frame directly on a whiteboard, and then time-adjusted photography techniques were applied to the final video. Most of these types of whiteboard videos combine stop-motion animation, bringing together sequential drawings in similar ways to traditional animation and flip books, with time-lapse photography to play back the recording of the stop-motion photography at a faster speed and then adding a voice-over to tell the story. The most common elements of whiteboard animation are the visualisation of a whiteboard as the canvas, the illustrator’s hand in the field of view, and a marker drawing out the story. With advancements in digital technology, these elements may be digital simulations that attempt to replicate the analogue. The representation of these elements is central to a rhetorical view of narrative in the ways that setting, voice and style are signified.

Although the medium is known by a variety of names that do not signal any sort of canvas (e.g. video scribing and fast draw video), as well as names that signal a different sort of canvas (e.g. sketchboard animation), whiteboard animation appears to be the dominant descriptor regardless of the true nature of the material practices. Since RSAAnimate videos gained popularity, other creative studios have found ways to simulate the appearance of this narrative form and their own art form as a version of whiteboard animation similar to the RSAAnimate in their menu of products. In fact, a Google search for RSAAnimate brings up links to a number of creative studios that replicate Cognitive Media’s whiteboard animation style; the official RSA site ranks fourth on the search results, with Whiteboard Animation, Idea Rocket Animation and Video Explainers showing up as the top three search results.

Though the final product may look similar to the audience, the process of whiteboard animation differs across studios and do-it-yourself (DIY) software tools. Sparksight combines old and new media by using hand-drawn illustrations, which are created on white paper with a black Sharpie marker and carefully photographed at a high shutter speed, then imported into an after-effects programme, spliced together and replayed at high speed to give the impression of movement (Sparksight Inc. 2015). Idea Rocket, on the other hand, takes a high-tech approach to simulate the analogue by using digital animation software for the illustration process, then compiling sketches into a movie, speeding up the film, adding complementary audio and, finally, integrating a green-screened hand as an additional video layer of the final movie (Idea Rocket 2017).

The production of these sorts of multimodal narratives are not limited to professional creative studios, as video scribing software and apps have made it possible for animators to participate in the evolution of digital animation and present stories in new ways. Though professional animators have been using computer input devices, such as graphic tablets with a stylus, to capture data and digitise it for years, the mass production of touch screen tablets coupled with animation software and apps allow ordinary people to draw on their digital devices and produce these sorts of narratives.

Critical issues and topics

With the visual and digital turn came an increase in adaptations of narrative forms that communicate in different ways. Desktop publishing, animation programmes, video editing, tools for digital illustration and social media allowed the ordinary person to become a self-published digital storyteller. In online media environments, YouTube provided space for producers and consumers of content to teach and learn from each another through video sharing and discussion spaces. With the surfacing of digital technologies and social media, user-generated content has proliferated with the rise of ‘prosumers,’ blurring the lines between amateurs and professionals (Toffler 1984). Tapscott and Williams (2010) later applied the term to describe the increased sharing of content and know-how between professionals and enthusiastic hobbyists that came with the emergence of Web 2.0 technologies. With these shifts, Jenkins (2007) suggests we are experiencing ‘the emergence of a new form of participatory culture (a contemporary version of folk culture) as consumers take media in their own hands, reworking its content to serve their personal and collective interests’ (para 5) such that the relationships between ordinary users, producers and content are more complex as old and new media converge. Innovation is happening from multiple directions, both top-down and bottom-up, as users experiment with a variety of techniques and devices for multimodal storytelling.

With cartoons serving as a primary tool to teach and engage people, this digital storytelling genre provides an example in which the mundane is engineered through the visual rhetoric of cartoons to be more appealing, and hence, persuasive. McCloud (1994) argued that comics and graphic novels are often thought of as simpler versions of their text-only counterparts, but should be considered as a significant storytelling form, deserving of respect. Similarly, Tiemensma (2009) pointed out that comics serve as an important medium for improving literacy as they combine printed words and pictures to represent narrative in today’s increasingly visual world. Drawing from cognitive literary theory, Nikolajeva (2014: 711) has shown how picture books are ‘a narrative art form that creates meaning through the synergy of two media, the verbal and the visual.’

Cartoons have a long history of teaching and captivating audiences. In the United States, political cartoons have been a vehicle to engage citizens in public discourse since the 1700s (Medhurst and Desousa 1981; Edwards 1999). In the 1970s, the animated musical series *School House Rock!* was used to teach children about grammar and civics, while adults were introduced to intellectual ideas via comic books in the *For Beginners* series. Educators have used cartoon-based video games to engage learning around serious topics such as the Middle Eastern crisis with *Darfur Is Dying* (Peng et al. 2010). The publication of the graphic doctoral dissertation *Unflattening* (Sousanis 2015) and a graphic special issue of *Digital Humanities Quarterly* (Whitson and Salter 2015) point to a significant shift in the level of acceptance for cartoons as an intellectually rigorous medium for teaching and presenting complex ideas through comic storytelling.

Main research methods

Through the lens of multimodality and rhetorical narratology, this chapter will investigate this little-studied form to reveal how technological transformations can inform storytelling practices, paying attention to material practices of production and multimodal narrative development.

Narrative

Early sociolinguistic research by Labov and Waletzky (1997) suggested that oral narratives of personal experience tend to have structural components (abstract, orientation, complicating action, resolution, evaluation and a coda) and could be understood through these formal frameworks. Scholars have interrogated traditional notions of narrative and call for research that develops frameworks for understanding the affordances that digital technologies bring to stories (Bamberg and Georgakopoulou 2008; Page 2009, 2012). Georgakopoulou (2006: 1) suggests that narrative research has ‘definitively moved from the study of narrative as text (first wave) to the study of narrative-in-context’ as the field enters a second wave.

This chapter reflects this second wave and explores narrativity and narrative discourse as social, rhetorical, and multimodal. Narrativity informs ‘almost any verbal utterance . . . ranging from fragmentary reports and abortive anecdotes to those more distinctly framed and conventionally marked tellings that we are inclined to call “tales” or “stories”’ (Smith 1980: 228). The social aspect of narrative discourse focuses on the act in which someone tells someone else something that happened, with each party motivated to participate in the telling of and listening to the narrative (Smith 1980; Bal 2004).

Narrative is not just about the genre, but about who tells the story and how it’s perceived. Rhetorical narratology has been most commonly applied to literary texts, although there is space to move beyond this. This narrative framework combines speech-act theory with rhetoric and narratology to consider the relationship between tellers and audiences as it occurs in specific contexts (Kearns 1999). Phelan (1996: 19) argued that a rhetorical approach to narrative ‘shifts emphasis from author as controller to the recursive relationships among authorial agency, textual phenomena, and reader response, to the way in which our attention to each of these elements both influences and can be influenced by the other two’.

Rhetorical approaches, regardless of their specific emphases, are ultimately concerned with the how, the what and the why of narrative (Phelan 1996). From a multimodal standpoint, narratives that are told through a combination of discursive resources that complicate meaning potentials, especially when resources are contradictory. This chapter considers the relationship between narrator and audience with consideration given to the multimodal dimensions of audio, visual and textual.

Multimodality

In combination with narrative theory, multimodal discourse studies provide an additional scaffolding to understand how whiteboard animation develops meaning through verbal and visual means. Semiotic resources carry their own small fragments of meaning-making that, when combined into a single presentation within an image, transmit a rhetorical message to the targeted audience (Jewitt et al. 2016). These fragments include design elements such as gestures, visuals and linguistic features within the image itself. Any combination of these elements builds a multimodal experience of meaning-making. Those who are the subject

of the communication are considered as represented participants and both interactive and represented participants shape the position of the viewer (Kress and van Leeuwen 2006). Multimodality decentres language and extends meaning making to a variety of modes of learning (Adami 2017; Jewitt 2013). Emphasising that meaning is not made in isolation from other modes of learning, multimodality challenges the primacy of verbal language (Cope and Kalantzis 2009; Adami 2017). That people express language from available resources or modes, such as writing, image, sound, speech, gesture, gaze and posture, and are active in its day-to-day production, is an underlying assumption of multimodality (Machin 2007).

Current contributions and research

Whiteboard animation is a multimodal storytelling form that is increasingly used to educate, inform, advertise and entertain. To uncover the persuasive power of digital storytelling in whiteboard animation, this chapter examines the complex relationship between speaker, audience, message and medium using a multimodal narrative analysis approach. Specifically, through a multimodal narrative framework, setting, time, point of view and voice will be analysed in commercial and DIY whiteboard animation examples.

Setting

Structure and text can provide a useful apparatus for considering the medium and mode of presenting a message – specifically comics animated on a whiteboard to tell a story. The rhetorical choice to situate animated stories on a whiteboard creates a context that influences message delivery. From a sociolinguistic perspective, the whiteboard can be seen as an orienting feature of the narrative (Labov and Waletzky 1997), providing a semiotic resource that signals the scene similar to the ‘setting’ in a literary narratives. Over time, the semiotic meaning of the whiteboard has shifted from primarily denotative to far more connotative in these animated stories.

Multimodal theorists point to the importance of props, such a whiteboard and marker, to impart meaning, especially in generic scenes (Machin 2007). At the beginning of each RSAAnimate video, for example, the series name is presented on a whiteboard. A hand holding a marker enters the field of view and writes the letters RSA on the white background while the sound of a marker squeaking across the slick surface of a whiteboard is audible with each stroke. This sensory information from multiple sources provides important cognitive information such that when the audience ‘sees’ a dry erase marker moving across a whiteboard when they ‘hear’ the squeaky marker. From a multimodal viewpoint, the use of sound complements the visual, such that the audio signals the white background as a specific type of space – a whiteboard.

With the literal visualisation of the whiteboard and thick marker denoting the scene, the symbolic meaning attached to the canvas imbues it with meaning potentials. To consider the possible connotative values associated with the whiteboard as an object, we may ask: where do we find whiteboards? what are whiteboards used for? and who uses whiteboards? Because ‘images are never free of connotative power’ (Machin 2007: 23), the visualisation of the whiteboard, be it authentic or synthetic, carries meaning that will be interpreted in specific ways. The associations of the medium as a note-taking space often used in classrooms and conference rooms to teach and share ideas carries connotative value such that the visual cues signal to the audience that the story produced will be educational and/or

informational. Moreover, the visualisation of the whiteboard context signals a virtual learning space, infusing the message with authority and credibility.

As the genre of ‘whiteboard’ animation has evolved, the representation of the whiteboard continues to serve as an important prop, but it has become more conceptual than physical, such that the actual whiteboard rarely exists – and if it does, it is merely a synthetic representation of it. The explicit move to simulate the whiteboard is present with the DIY app Doceri, which is described as ‘the interactive whiteboard for iPad’ (SP Controls, Inc. 2015). While Doceri provides animators with a variety of background options, including a chalkboard or blank background (in customisable colours), the default setting for background is a slide of a whiteboard, complete with the aluminium frame around the edges. For some creative studios, the whiteboard is implied through complementary tools and previous videos that have normalised an understanding of the genre as ‘whiteboard animation.’ Sparksight uses old school techniques of paper, Sharpie markers and stop-motion photography, whereas Idea Rocket uses a white background with digital illustration software. The whiteboard is now taken for granted as the composing canvas, and viewers no longer need to see an authentic version. The final video produced with these tools is understood by the audience as a whiteboard video regardless of whether a whiteboard exists.

Time

Temporality is another orientating device that is inextricably tied with narrative (Ricoeur 1984). Multimodal approaches provide tools for critiquing types of images and presentation in terms of meaning potentials, a multimodal narrative approach allows for consideration of sequencing and temporality in the creation process. With whiteboard animation, time is managed through the telling and visualisation of events in a story. As the audience listens to the narrator and watches the animator draw on whiteboard, the story develops in front of the viewer’s eyes. With each marker stroke on the canvas, the audience anticipates story elements as the animator takes the viewer through the narrative.

While temporal sequencing is integral to narrative, linear disruption is commonly employed for effect. For example, the narrator may refer to events out of order or return to an earlier time in a story to add details. In literary narrative, authors may do this by presenting story scenes out of order, such as presenting the climax of the story as an opening scene, or through reading order, as seen with books like the ‘choose your adventure’ franchise of the 1980s and the digital novel *Arcadia* (Pears 2015), both of which allow the audience to drive the sequencing of events via choices about which path to follow in the reading process. With whiteboard animation, linear disruption may be realised both through the telling of the story and/or through the animation process – and not always simultaneously. For example, while the story may be told with a forward moving linear progression, the visualisation of the story may be recursive when a visual story element is brought back into the frame, bringing previously presented illustrations back into the audience’s field of view, and the illustration is revised.

Ewert (2010: 72) has pointed out that narratology involving comics must consider textual and graphic elements, which is often ‘a challenge for critics habituated to text-based narrative’. Narrative time is manipulated in comics through visual space in a variety of ways (Eisner 2008). For example, animators may integrate symbolic resources that connote time (like images of a clock), or they may use space between and within panels to represent time as compressed or extended by using more or less white space. The concept of ‘closure’ describes the phenomena when a reader is able to imagine the action that has taken place

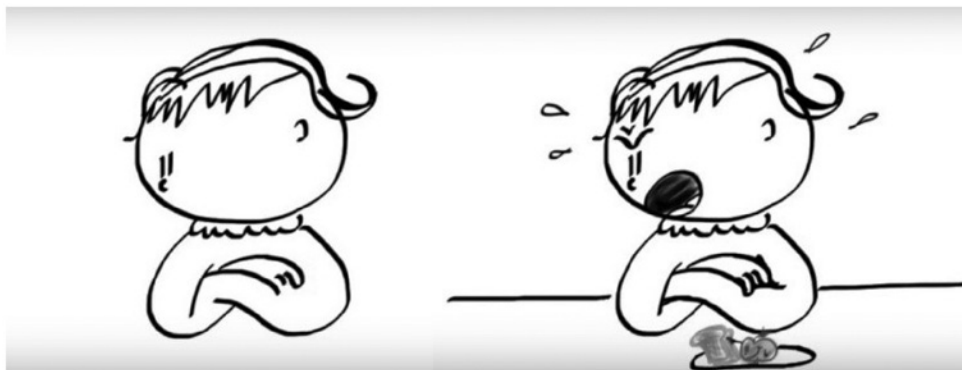


Figure 24.1 *Have You Met Olivia?*

Source: Childhaven 2014

between drawn panels and ‘close the gap’ about what is expected to happen in the blank space between panels (McCloud 1994).

The concept of closure is visible in the DIY whiteboard animation *Have You Met Olivia?* (Childhaven 2014). In a scene that depicts the child as a picky eater, the audience watches as the pen strokes draw Olivia with arms crossed. Initially, the face includes only eyes and a nose, leaving the face virtually expressionless. Next, a plate and food items are drawn into the sequence as if to present the audience with the conflict that the main character must overcome – eating the food in front of her. Finally, as the third-person narrator describes how Olivia had outbursts when food was placed in front of her, the main character’s mood shifts from an emotionless state to anger as the animator fills in the details of the story one pen stroke at a time. Initially, the only indicator that the audience has of Olivia’s mood is relayed through her crossed arms. It isn’t until the animator draws a screaming mouth, furrowed eyebrows and beads of sweat flying into the air that the audience learns of the extent of Olivia’s upset, revealing the narrative’s conflict (Figure 24.1). In this way, stroke order serves as a temporal narrative technique, with each stroke of the marker foreshadowing the completed image to come.

Point of view: who sees the story

It has been suggested that

few narrative features have been discussed as extensively as point of view – the physical, psychological, and ideological position in terms of which narrated situations and events are presented, the perspective through which they are filtered – and few have been associated with as rich a terminology (from central intelligence, vision, and focalization to filter and slant).

(Prince 2010: 442)

Point of view (POV) describes the position of the narrator in relation to the story being told and provides the audience with a perspective on events, which can emanate from a character or a nonanthropomorphic existence such as a camera. In this way, POV is the lens through

which an audience is influenced to ‘see’ how events of a story unfold. Thus, POV is a narrative device used to give readers insight into events in the story and is achieved differently depending on the media.

In literary genres and oral narratives, linguistic devices are the tools that provide perspective. First person and third person are the most common point of view techniques used in literary narratives, with pronouns doing the work of signalling to the audience. In first person, the narrator tells the audience how ‘I’ progressed through the story, whereas with third person, all characters are spoken of in an objective manner through pronouns such as he, she and they. The lesser-used second-person POV, in which the narrator speaks directly to ‘you,’ the audience, tends to be found most often in self-help, instructional and technical communication story forms.

POV is an important tool for negotiating social distance between the narrator and audience. Because first-person narrative provides a window into the main character’s mind as they narrate the story through their POV, it closes social distance by giving readers a sense that they know the narrator and are with them on the narrative journey. While third person does give the audience access to the narrator’s mind and keeps them close as the narrator tells the audience what they want them to know at any given moment, first person puts the audience directly into the character’s mind, unlike third person which moves the audience out of it. With second-person POV, the audience does not have access to the narrator’s thoughts, yet the direct address works to include them into the narrative.

In film and video games, POV is achieved through camera angles that allow the audience to see events through the perspective of particular characters. With first person, the audience sees events through the eyes of the main character as represented via the camera lens. In third person, the audience still sees what the main character sees, but the audience does not have a window into the main character’s thoughts as the camera appears to be looking over the shoulder of, or even including, the main character in the field of view. Second-person POV is used when the narrator speaks directly to the camera and uses the pronoun ‘you’ to address the audience. POV can refer to the visualisation of what a character sees, or more abstractly, as used with narratology approaches, POV can reference attitudes and tendencies of the cinematic narrator (Keating 2010).

With whiteboard videos, POV is complicated by the multimodality of the story form as layers of textual and cinematic narration are combined to tell a story. First, there is the spoken narrative that the audience hears as the lecturer talks to the audience, the perspective of the camera that films the animator drawing the story and, finally, the POV is presented through cartooning techniques. This is not to say that POV is the same across whiteboard animation. Rather, POV in these narratives often reflects stylistic diversity through software and animation tools used to tell the story.

As creative studios have experimented with this storytelling form, shifts in POV can be seen even within a set of seemingly similar narratives. Across the 21 RSAnimates created between December 2009 and July 2016, visual POV varies. While the verbal narrative POV has largely remained constant, with narrators switching between first-, second-, and third-person perspective throughout the course of their lecture, the visual perspective has evolved with technological and cinematic shifts.

RSAnimate provides an example of how POV can be used to push the audience away or bring them closer to the story through camera angles. Early whiteboard animations primarily used over-the-shoulder camera angles with periodic medium long shots. Visually, the use of close-up over the shoulder camera angle gives the audience a first-person perspective as they watch the animator’s thought process with each stroke drawn. As the camera angles

move from the over-the-shoulder to the medium long shot, the camera captures the animator's shoulder, neck and back of the head as well as the back of the body from the knees up, giving the audience an opportunity to step back and watch the animator telling the visual story. Thus, the camera utilises a first- and third-person visual perspective, showing the story through the eyes of the animator, as well as stepping away to show the animator. As the RSAAnimate series evolved, the medium long shot was abandoned and only the over-the-shoulder and close-up perspective remained, with only the drawing hand marker strokes visible. With this evolution of perspective, in which the camera lens moves into to focus on the drawing hand and the animation in process, the boundary between the animator and the speaking narrator becomes blurred.

An underlying assumption of multimodality is that people express language from available resources or modes such as writing, image, moving image, sound, speech, gesture, gaze and posture (Jewitt 2013: 253). Image types and photographic perspectives interact with the audience to perform visual speech acts and simulate social interaction. Body parts, such as arms, legs and fingers, or held objects, like hand tools, can be used to establish connectivity, as can the gaze that visually addresses and demands an action of the viewer (Kress and Leeuwen 2006). Moreover, perspective angles within images can establish vector patterns between semiotic resources. Specifically, a vector pattern is an interaction that '[does] something to or for each other', such as a character pointing a finger at another person or object (Kress and Leeuwen 2006: 59). For the animator visualising the narrative, the shifts in POV need to be managed through cartooning techniques that prompt interaction between characters and the audience. In cartooning, visual 'demands' that simulate social interaction may include gestures, eye contact and speech bubbles in which the character directly addresses the audience.

In *RSA: Changing the Education Paradigm* (2010a), POV is realised linguistically through the use of the first-person pronoun 'we' and second-person pronoun 'you' and visualised through strategic integration of a caricature of the RSA lecturer Sir Ken Robinson. In responding to his own question of whether we should keep academic standards high, the caricature of Robinson is drawn facing the audience and speaking via the cartooning device of the speech bubble. The caricature of the narrator points his finger in the air to accentuate his point and speech bubbles represent the linguistic engagement moves; Robinson uses a solidarity move by saying, 'Yes! We should!' and then asks the audience 'why would you lower [standards]?' speaking directly to the audience while making eye contact. In a follow-up frame, the caricature of Robinson points directly at the audience while telling them, 'If you are interested in the model of education, you don't start from production line mentality'. In these examples, POV is enacted on both linguistic and semiotic levels, with second-person pronouns, speech bubbles, the demand gaze and pointing gestures to draw the audience into the story (Figure 24.2).

Similarly, in *RSA: Economics Is for Everyone!* (2016), the narrator, RSA lecturer Ha-Joon Chang, is also integrated into the animation to shift POV. When Chang linguistically shifts between first-person, second-person and third person POV in his lecture, it is also visually captured through speech bubbles and hand gestures. As Chang uses second-person pronouns to talk to the audience, saying 'This is the attitude you have to have in debating economic issues', visual perspective is achieved through caricatures of Chang gesturing with an open hand and looking at the audience. In another frame, where the narrator attempts to create solidarity by stating that 'We all need these diverse approaches to economics', the pronoun 'we' is operationalised along with the demand gaze and outreached arms that invite the audience to be part of the larger group. Additionally, special effects are used to show Chang blinking, making the cartoon version of the narrator seem more real and



Figure 24.2 RSanimate – Changing the Education Paradigm (2010a)

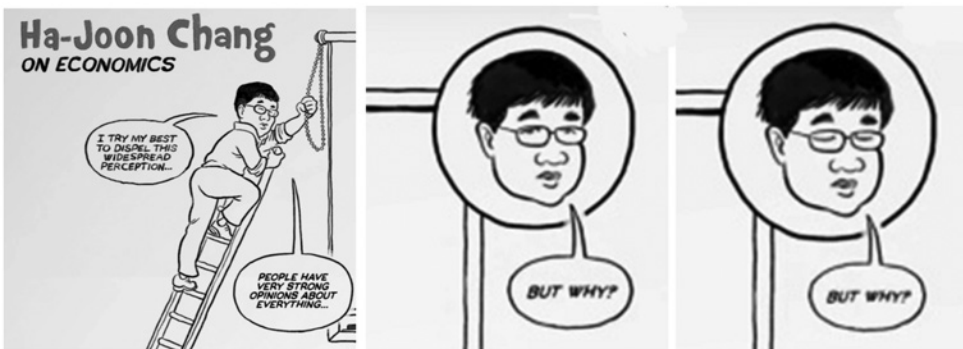


Figure 24.3 RSanimate – Economics Is for Everyone! (2016)

rhetorically amplifying the demand gaze (Figure 24.3). This move to enhance authenticity increases modality and, following social semiotic theory, images that act with higher modality are deemed more credible and hold more persuasive power with audiences.

One way in which digital composers can appeal to the emotions is by manufacturing a sense of intimacy between the narrator and the audience. This sort of synthetic personalisation can be seen when a message for a mass audience is engineered in such a way that individuals feel as if they are personally addressed (Fairclough 1996). By employing visual resources to befriend an online audience through linguistic and visual speech acts that rhetorically manufacture a sense of intimacy and closeness, digital composers engage in the business of *virtual visual synthetic personalisation* to sell ideas and products (Thompson 2012). Multimodality is helpful for making sense of rhetorical appeals that bring together a complex web of audio and visual modes for communicating casually with online audiences. In the RSanimate series, caricatures of narrators speak and gesture directly to the audience, with camera angles allowing the audience a sense that they are sitting with the animator and watching the narrative process from the same perspective – these are just some of the ways in which whiteboard animators manipulate POV to connect with the audience. While the story is narrated by the lecturer with an audio voice-over, it is simultaneously told through a cartoon caricature of the lecturer. Synthetic personalisation, be it verbal or visual, is a rhetorical device that appeals to the emotions of the audience (pathos). Aristotle (2004: 9) pointed to such appeals as one of the most important tools of persuasion such that ‘the emotions are all those feelings that so change men as to affect their judgements, and that are also attended by pain or pleasure’.

Voice: who tells the story

POV and voice are narrative devices that are often conflated. The best way to differentiate these elements is to think of POV as ‘who sees the story’ and voice as ‘who tells the story’. This is not unlike Smith and Watson’s (2010) discussion of autobiographical ‘I’ as multiple with a distinction between the narrator and the narrated. In whiteboard animation, these devices of authorial control can be complicated by the levels of narration and POV that work on the audience simultaneously. There is the narrator telling the story verbally, as well as an animating narrator who tells the story visually, so the audience ‘sees’ the visual voice. Thus, a speaker provides one level of voice and an animator adds another, making whiteboard animation rich with multivocality. In literature, multiple narrators are employed to represent a story from different perspectives with distinct voices. In whiteboard animation, however, the voices of the narrator and animator are combined to create the impression that they are one and the same. This is further complicated when a caricature of a narrator is drawn into the story to serve as an orator who speaks to the audience, such that authorial control can be a combination of the speaking narrator, an illustrating cartoonist (via some form of embodiment) and a cartoon version of the narrator telling the story. By taking a multimodal narrative approach, these layers of voice can be examined.

As whiteboard animation has evolved, the drawing hand has come to represent a significant element of authorial control. In the earliest RSAnimates, camera angle alternated between mid-shot and close-ups of the back of the animator and his drawing hand. A comparison between early and recent videos demonstrates a clear move from a process in which the animator is recorded drawing on a sketchboard to a technologically simulated version of a hand animating on a whiteboard. As the RSAnimate series has evolved, camera angles have moved in and the animator’s presence has moved out with only the drawing hand and forearm visible through close-up shots. In the most recent RSAnimates, the sporadic integration of the hand appears to be a visual effect of animation software that marks the genre rather than an authentic representation of the hand in motion. With the new process, the body of the animator is no longer present, and the hand is only sporadically visible (if at all). For RSAnimate, the visual narrator has slowly been erased, with only the marker strokes left to follow as the story unfolds. In a tutorial about the process used by Sparksight, the animator discusses the importance of the hand and how after-effects editing is used to make the drawing hand appear uninterrupted and constant (Sparksight 2013).

From a rhetorical narratology approach, ‘voice is the fusion of style, tone, and values’ that link to attitudes, ways of talking and ideology for the audience (Phelan 1996: 45). An analysis of RSAnimate shows that audio voices narrating the series are overwhelming male, at 86 percent, and while 14 percent of RSA lectures are given by women, all the corresponding whiteboard animation versions are visually voiced through what RSA describes as ‘everyone’s favourite hairy hand’ on their website (Figure 24.4). Phelan (1996: 43) has pointed out that ‘To listen to the narrative is, in part, to listen to values associated with a given way of talking’, such that the visual re-voicing of female narrators means that the audience only sees the narrative through the gendered male body, problematically eliding the voices of female intellectuals. Since there has been a historical lack of representation of women in academia (Winslow and Davis 2016) and comics (McCloud 2000), this sort of visual re-voicing through the animation reinforces the problem of gender inequity in subtle ways. One may apply Bakhtin (1984) conceptions of ‘polyphony’ (where more than one voice can be heard) and ‘double-voicing’ (where a different voice is adopted to the discourse of another) to whiteboard animation to understand how narrative voice can be ‘overpopulated’ with the



Figure 24.4 RSAnimates: *Smile or Die* (2010b); *The Paradox of Choice* (2011); *How to Help Every Child Reach Their Potential* (2015)

voices of others (Aczel 2010b). And while the RSAnimate examples of double-voicing can be described as ‘uni-directional’ in that the goal is to maintain the original discourse through the stylisation of animation (Aczel 2010a), the stylistic choice to re-gender the animator, or ignore the option to maintain the original gender, politicises multimodal narrative.

The inclusion of the hand, as well as a specific type of hand, is a rhetorical choice. DIY programmes like Sketch Book Pro 3.0 allow composers to choose from drawing hands or even import their own hand. Though the ‘hand’ library for VideoScribe is limited to pre-set categories of gender (male and female), age (adult or child) and race represented by colour (white, brown and black), it is progressive in representing the embodied narrator voice in diverse ways. DIY users have called for more diverse choices in the hand library on the help forum and VideoScribe has responded with additions. However, at the time of this writing, there was only one black-skinned hand. Thus, the hand that is a hallmark of the genre is rarely the true embodiment of the illustrator, but rather has become a digital element to be manipulated and customised for stylistic effect, with the hand serving as a rhetorical device that visually constructs the animator based on a limited number of options. Similarly, software programmes often allow composers to choose the writing implement that will be shown in the drawing hand; options include a pen, pencil, crayon, paint brush and, of course, a whiteboard marker. In this way, the combination of the hand and the writing tool represent an aspect of authorial control that works in concert with the voice of the narrator and the visual imagery of a cartoon character talking to the audience.

Future directions

With the continuing emergence of new forms of digital storytelling, methods that move beyond linguistic tools of narrative analysis are needed to adequately examine these complex multimodal narrative forms. As the tools of production for digital composing continue to become more readily available and user-friendly, avenues for distribution and consumption become more accessible, making research that takes a critical approach to these elements vital. In the last decade, DIY software programmes have emerged allowing ordinary people to tell their stories through whiteboard animation through libraries that include pre-drawn characters, stock images, music, drawing implements and hands. Just as noted in this chapter, future research must continue to examine issues of representation in professional and DIY whiteboard animation.

As the genre of whiteboard animation has evolved, the ways in which narrative elements work on the audience has also shifted. The current use of the whiteboard as a canvas

carries with it connotations specific to material practices associated with the writing context, whereas innovations to this context have the potential to create new ‘settings’ for audiences to imagine in the future. Narrative time can be managed multimodally through audio narration, visual space and stroke order, as shown in these examples, but also through devices such as music, iconic cartoon devices and after-effects in the post-production phase.

Research that examines whiteboard animation through the lens of multimodal narrative fits within a Big Tent DH framework that looks to understand the intersection between technology and storytelling. At the annual *Day in the Life of the Digital Humanities*, Selber argued that DH is ‘the study of the role of technology in helping to structure the work of writers and writing contexts’ (TAPoR 2011). With a similar concern for rhetorical forms, Piez (2008: 9) called for ‘media consciousness’ and pointed out that ‘not only do we study digital media and the cultures and cultural impacts of digital media; also we are concerned with designing and making them’. Thus, future research that considers other aspects of multimodal narrative devices would also contribute to DH scholarship that looks to understand the implications of technology, writing and narrative.

Further reading

- 1 Comer, K. (2015). Illustrating praxis: Comic composition, narrative rhetoric, and critical multiliteracies. *Composition Studies* 43(1): 75–104.
This essay proposes using narrative theory alongside comic studies, with a focus on narrative gaps, narration and focalisation as a heuristic for rhetorical design to foster critical multiliteracies.
- 2 Machin, D. (2013). What is multimodal critical discourse studies? *Critical Discourse Studies* 10(4): 347–355.
This introduction to a special issue of *Critical Discourse Studies* looks to bring a critical lens to multimodal discourse studies, illustrating how different levels of communicative activity are infused by, and shaped by, power relations and ideologies through semiotic resources.
- 3 Royce, T.D. (2015). Intersemiotic complementarity in legal cartoons: An ideational multimodal analysis. *International Journal for the Semiotics of Law – Revue Internationale de Sémiotique Juridique* 28(4): 719–774.
This article applies multimodal discourse analysis to legal cartoons and argues to extend the framework to address more complex layers of represented semiotic meaning.
- 4 Souto-Manning, M. (2014). Critical narrative analysis: The interplay of critical discourse and narrative analyses. *International Journal of Qualitative Studies in Education* 27(2): 159–180.
This article looks to establish critical narrative analysis by bringing together critical discourse analysis and narrative analysis to address theoretical and methodological dilemmas to examine power inequities in narratives.

References

- Aczel, R. (2010a). Polyphony. In D. Herman, M. Jahn and M.-L. Ryan (eds.), *Routledge encyclopedia of narrative theory*. London: Routledge.
- Aczel, R. (2010b). Voice. In D. Herman, M. Jahn and M.-L. Ryan (eds.), *Routledge encyclopedia of narrative theory*. London: Routledge.
- Adami, E. (2017). Multimodality. In *The Oxford handbook of language and society*. Oxford: Oxford University Press.
- Aristotle. (2004). *Rhetoric* (D. Thrift, eds.). New York: Dover Publications.
- Bakhtin, M. (1984). *Problems of Dostoevsky's poetics*, 1st ed. Minneapolis: University of Minnesota Press.
- Bal, M. (2004.) *Narrative theory: Major issues in narrative theory*. Taylor & Francis.

- Bamberg, M. and Georgakopoulou, A. (2008). Small stories as a new perspective in narrative and identity analysis. *Text & Talk* 28(3): 377–396.
- Borek, L., Dombrowski, Q., Perkins, J. and Schöch, C. (2016). TaDiRAH: A case study in pragmatic classification. *Digital Humanities Quarterly* 10(1).
- Childhaven. (2014). *Have you met Olivia?* Available at: www.youtube.com/watch?v=GhrIYJ_s6Xw (Accessed 5 March 2017).
- Cognitive Media. (2017). *Explainer videos & digital storytelling – previous work*. Available at: www.wearecognitive.com/our-work (Accessed 18 February 2017).
- Comer, K. (2015). Illustrating praxis: Comic composition, narrative rhetoric, and critical multiliteracies. *Composition Studies* 43(1): 75–104.
- Cope, B. and Kalantzis, M. (2009). “Multiliteracies”: New literacies, new learning. *Pedagogies: An International Journal* 4(3): 164–195. doi:10.1080/15544800903076044.
- Edwards, R. (1999). Politics as social history: Political cartoons in the gilded age. *OAH Magazine of History* 13(4): 11–15.
- Eisner, W. (2008). *Comics and sequential art: Principles and practices from the legendary cartoonist*. New York: W. W. Norton & Company.
- Ewert, J.C. (2010). Comics and graphic novel. In D. Herman et al. (eds.), *Routledge encyclopedia of narrative theory*. London: Routledge.
- Fails, J.A. and Kelton, A. (2016). Call for papers. *The Journal of Emerging Learning Design*. Special Issue: *The Digital Humanities*. Available at: www.hastac.org/opportunities/call-papers-journal-emerging-learning-design-special-issue-digital-humanities (Accessed 13 February 2017).
- Fairclough, N. (1996). Technologicalisation of discourse. In C.R. Caldas-Coulthard and M. Coulthard (eds.), *Texts and practices: Readings in critical discourse analysis*. London: Routledge, pp. 71–83.
- Georgakopoulou, A. (2006). Thinking big with small stories in narrative and identity analysis. *Narrative Inquiry* 16(1): 122–130. doi:10.1075/ni.16.1.16geo.
- Idea Rocket. (2017). *IdeaRocket – animated video production services: 2d, 3d, whiteboard*. Available at: <http://idearocketanimation.com/> (Accessed 19 February 2017).
- Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. New York: New York University Press.
- Jenkins, H. (2007). From participatory culture to participatory democracy (Part Two). *Confessions of an Aca-Fan*. Available at: http://henryjenkins.org/blog/2007/03/from_participatory_culture_t_1.html (Accessed 21 February 2018).
- Jewitt, C. (2013). Multimodal methods for researching digital technology. In S. Price et al. (eds.), *The Sage handbook of digital technology research*. Los Angeles: Sage Publications, pp. 250–265.
- Jewitt, C., Bezemer, J. and O’Halloran, K. (2016). *Introducing multimodality*. London: Routledge.
- Kearns, M.S. (1999). *Rhetorical narratology*. Lincoln, NE: University of Nebraska Press.
- Keating, P. (2010). Point of view (Cinematic). In D. Herman et al. (eds.), *Routledge encyclopedia of narrative theory*. New York: Routledge.
- Kress, G.R. (2009). *Multimodality: Exploring contemporary methods of communication*. Abingdon and New York: Routledge.
- Kress, G.R. and Leeuwen, T. van (2001). *Multimodal discourse*. London and New York: Bloomsbury Academic.
- Kress, G.R. and Leeuwen, T.V. (2006). *Reading images: The grammar of visual design*. Abingdon: Routledge.
- Labov, W. and Waletzky, J. (1997). Narrative analysis: Oral versions of personal experience. *Journal of Narrative and Life History* 7(1–4): 3–38.
- Machin, D. (2007). *Introduction to multimodal analysis*. New York: Oxford University Press.
- Machin, D. (2013). What is multimodal critical discourse studies? *Critical Discourse Studies* 10(4): 347–355. doi:10.1080/17405904.2013.813770.
- McCloud, S. (1994). *Understanding comics: The invisible art*, reprint ed. New York: William Morrow Paperbacks.

- McCloud, S. (2000). *Reinventing comics: How imagination and technology are revolutionizing an art form*. New York: William Morrow Paperbacks.
- Medhurst, M.J. and Desousa, M.A. (1981). Political cartoons as rhetorical form: A taxonomy of graphic discourse. *Communication Monographs* 48(3): 197–236. doi:10.1080/03637758109376059.
- Nikolajeva, M. (2014). Picturebooks and emotional literacy. *The Reading Teacher* 67(4): 249–254. doi:10.1002/trtr.1229.
- Page, R. (2009). *New perspectives on narrative and multimodality*. New York: Routledge.
- Page, R. (2012). *Stories and social media: Identities and interaction*. New York: Routledge. Available at: www.routledge.com/books/details/9780415889810/ (Accessed 19 April 2014).
- Pears, I. (2015). *Arcadia: A novel*. New York: Knopf.
- Peng, W., Lee, M. and Heeter, C. (2010). Les effets d'un jeu sérieux sur la prise de rôle et la volonté d'aider. *Journal of Communication* 60(4): 723–742. <https://doi.org/10.1111/j.1460-2466.2010.01511.x>
- Phelan, J. (1996.) *Narrative as rhetoric: Technique, audiences, ethics, ideology*. Columbus: Ohio State University Press.
- Piez, W. (2008). Something called digital humanities. *Digital Humanities Quarterly* 2(1).
- Prince, G. (2010). Point of view (Literary). In D. Herman et al. (eds.), *Routledge encyclopedia of narrative theory*. London: Routledge.
- Ricoeur, P. (1984). *Time and narrative*, vol. 1. Chicago, IL: University of Chicago Press.
- Royce, T.D. (2015). Intersemiotic complementarity in legal cartoons: An ideational multimodal analysis. *International Journal for the Semiotics of Law – Revue internationale de Sémiotique juridique* 28(4): 719–744. doi:10.1007/s11196-015-9421-1.
- RSA: The royal society for the encouragement of arts, manufactures and commerce. (2010a). *Changing education paradigms: Sir Ken Robinson*. Available at: www.youtube.com/watch?v=zDZFcDGpL4U (Accessed 17 February 2017).
- RSA: The royal society for the encouragement of arts, manufactures and commerce. (2010b). *Smile or die: Barbara Ehrenreich*. Available at: www.thersa.org/discover/videos/rsa-animate/2010/03/rsa-animate-smile-or-die (Accessed 17 February 2017).
- RSA: The royal society for the encouragement of arts, manufactures and commerce. (2016). *Economics is for everyone!* Available at: www.youtube.com/watch?v=NdbbcO35arw (Accessed 2 February 2017).
- RSA: The royal society for the encouragement of arts, manufactures and commerce. (2011). *The paradox of choice: Renata Salecl*. Available at: www.thersa.org/discover/videos/rsa-animate/2011/06/rsa-animate-choice (Accessed 17 February 2017).
- RSA: The royal society for the encouragement of arts, manufactures and commerce. (2015). *How to help every child fulfil their potential: Carol Dweck*. Available at: www.thersa.org/discover/videos/rsa-animate/2015/how-to-help-every-child-fulfil-their-potential (Accessed 24 July 2018).
- Smith, B.H. (1980). Narrative versions, narrative theories. *Critical Inquiry* 7(1): 213–236.
- Smith, S. and Watson, J. (2010). *Reading autobiography: A guide for interpreting life narratives*. Minneapolis: University of Minnesota Press.
- Sousanis, N. (2015). *Unflattening*. Cambridge, MA: Harvard University Press.
- Souto-Manning, M. (2014). Critical narrative analysis: The interplay of critical discourse and narrative analyses. *International Journal of Qualitative Studies in Education* 27(2): 159–180. doi:10.1080/09518398.2012.737046.
- Sparksight. (2013). *How to do a fast draw animated video*. Available at: www.youtube.com/watch?v=QvyziGZJGM0 (Accessed 3 March 2017).
- Sparksight Inc. (2015). Corporate video: Animation vs. live action. *Sparksight Blog*, 22 September. Available at: <https://sparksight.com/blog/corporate-video-animation-vs-live-action/> (Accessed 22 April 2018).
- SP Controls, Inc. (2015). *Doceri – the interactive whiteboard for iPad*. Available at: <https://doceri.com/> (Accessed 3 March 2017).
- TAPoR. (2011). *How do you define humanities computing/digital humanities?* Available at: www.artsrn.ualberta.ca/taporwiki/index.php/How_do_you_define_Humanities_Computing/_Digital_Humanities%3F (Accessed 21 July 2016).

- Tapscott, D. and Williams, A.D. (2010). *Wikinomics: How mass collaboration changes everything*, expanded ed. New York: Portfolio.
- Terras, M. (2011). Peering inside the big tent: Digital humanities and the crisis of inclusion. *Melissa Terras' Blog*, 26 July. Available at: <http://melissaterras.blogspot.com/2011/07/peering-inside-big-tent-digital.html> (Accessed 16 July 2018).
- Thompson, R. (2012). Looking healthy: Visualizing mental health and illness online. *Visual Communication* 11(4): 395–420. doi:10.1177/1470357212453978.
- Tiemensma, L. (2009). Visual literacy: To comics or not to comics? Promoting literacy using comics. In *World library and information congress: 75th IFLA general conference and council*. Milan, Italy. Available at: www.ifla.org/past-wlic/2009/94-tiemensma-en.pdf (Accessed 1 June 2016).
- Toffler, A. (1984). *The third wave*. New York: Bantam.
- Whitson, R.T. and Salter, A. (2015). Introduction. *Comics and the Digital Humanities* 9(4). Available at: www.digitalhumanities.org/dhq/vol/9/4/000210/000210.html (Accessed 28 February 2017).
- Winslow, S. and Davis, S.N. (2016). Gender inequality across the academic life course. *Sociology Compass* 10(5): 404–416. doi:10.1111/soc4.12372.

English language and digital literacies

Paul Spence

Introduction

It is widely accepted that digitally mediated culture and technology have had a significant impact on human society on various levels in the last two decades, but it is less clear what impact this has, or might have, in the humanities.¹ Various responses have attempted to situate digital culture, both theoretically and practically, in relation to research in these disciplines – from new media and digital humanities perspectives (Svensson and Goldberg 2015 or Liu 2011), among others – but rather less attention has been paid to implications for the process of learning. What are the intellectual and practical skills required for learners to master these new tools? How should we use digital methods to study culture, society and the human record, and how might we integrate them into our critical toolkit? What impact does the proliferation of media, code and data have on how we work with source materials; how do the resulting *remediations* (Bolter and Grusin 2000) in digital form influence our analysis and interpretation; and what are the consequences for curricula in the humanities?

This is not a new debate. As far back as 1988 Cynthia Selfe was reminding us that computers have been part of the apparatus of scholarship for some time, and she argued that the battles over the ‘computer revolution’ in the English classroom have not ended yet. We are, she proposed, ‘just starting to consider the implications of the multi-layered literacy associated with computers’ (Selfe 1988: 55, cited in Kirschenbaum 2010). In the three decades since this was published, there is not much evidence that this debate has shifted significantly in a programmatic manner, notwithstanding some pioneering projects and innovative educational experiments, many of them connected to new media studies or the digital humanities. This is surprising when we consider how much has changed in the general conditions in which this debate takes place. It is more common to hear talk of ‘digital culture’ rather than the narrower focus on technical expressions such as ‘technology’ or ‘computer’. Digital culture has permeated most of our lives in ways that could not have been foretold back then. And there is, it is true, a greater sense that this is a social process rather than just a question of technology and practical skills. But in spite of all of these factors it is still far from clear how digital culture and technology affect the study of English (or the humanities more generally) in the twenty-first century.

This chapter examines these challenges through the term ‘digital literacy’ and its variants. While the term’s meaning and application are somewhat ambiguous and contested, the very nature of this variation provides useful material for discussion about different, and often conflicting, attitudes toward the process of learning in digitally mediated environments. I argue here that this encourages us to re-examine some of the prevailing assumptions about the role of ‘the digital’ in humanities education.

Definitions

Lankshear and Knobel observe that ‘literacy’ is now a core part of our understanding of education in a way which would not have been imaginable a few decades ago, when there was a greater emphasis on the concept of ‘reading’ (Lankshear and Knobel 2003: 3), a transition they attribute to numerous factors, including the influence of Paulo Freire’s ‘radical education movement’, educational programmes related to literacy drives in the United States in the 1970s and an increasing sociocultural slant to education, which elevated the concept in academic terms and led to a well-funded literacy ‘industry’ (Lankshear and Knobel 2003: 5–7). Historically, the term ‘literate’ has its roots in the concept of ‘being learned’ and an appreciation and mastery of literature, but the basis for our current understanding of ‘literacy’ tends to be framed in the modern era around the concepts of reading and writing (sometimes with a close association to numeracy), and this fact still strongly influences current usage in many educational contexts: see, for example its use in the UK National Curriculum² or the UK Literacy Association.³ In recent years the term has evolved considerably to broader definitions, often stressing connection to social and economic development, and sometimes simply referring to ‘competence’ or to the ability to ‘find one’s way around some kind of system’ (Lankshear and Knobel 2003: 14–15). A UNESCO report in 2006 offered four general labels which claimed to capture most interpretations of the term: (1) literacy as text or discourse; (2) literacy as applied, practised and situated; (3) literacy as an autonomous set of skills; and (4) literacy as a learning process (UNESCO 2006).

In most of these formulations, literacy is no longer viewed as an individual experience or as a homogenous concept, but something which is social, situated and context dependent. Current concepts of ‘literacy’ are in part influenced by the work of ‘new literacy studies’ scholars in the late twentieth century, who looked to redefine ‘literacy’ in sociocultural terms, downplaying its traditional cognitive emphasis, and the broader concept of ‘new literacies’ proposed by Lankshear and Knobel and others, which slightly later acknowledge the impact of computer and Internet-mediated practices on learning (Lankshear and Knobel 2003). Many of the commonly used expressions which include the word ‘literacy’ – terms such as ‘information literacy’, ‘visual literacy’, ‘spatial literacy’ or ‘media literacy’ – point to the need for a richer understanding of human communication in today’s society, beyond the traditional focus on text and the printed word, and which encompass a wider set of media and content types (images, audio-visual material, maps), which are often linked to digitally related phenomena such as information abundance and media saturation.

The concept of ‘digital literacy’, or its plural form ‘digital literacies’, operates in this context of ambiguous and dynamic definition, and has interpretations which are located across a spectrum ranging from the view of ‘digital’ as mere enhancement of traditional education (learning a particular technical skill) to more critical constructions of the term, which emphasise the need to engage with a new set of social and communicative processes arising from the ‘digital turn’. I favour the plural form ‘digital literacies’ here (unless referring to

usage by others) to reflect the multiplicity of meaning behind the term and to connect with the numerous different learning practices which they entail.

Background

Digital culture and technology have penetrated human culture and society in multiple interlocking ways. It is estimated that global Internet access is currently at 49.6 percent,⁴ reaching 97 percent in some parts of the world.⁵ A recent Reuters survey found that 51 percent of people in the countries surveyed used social media (such as Twitter or Facebook) as a source for their news about the world,⁶ and in some countries, a high percentage of school children have access to mobile telephones, many of them from a young age (Clark and Douglas 2011). This sets humans up for a radically different encounter with many aspects of culture, which are now increasingly digitally mediated in the spheres of content production, analysis and publication. While the scope and depth of this process are profoundly uneven, this has huge implications for how we engage with the human record and for how knowledge is created and disseminated.

Digital literacies

The concept of ‘digital literacies’ has developed as a result of the interaction between evolving educational debates about pedagogy and waves of digital innovation, themselves socially inflected (Web 2.0, mobile communication and open versus ‘walled garden’ device/content models). The term ‘digital literacy’ gathered steam in the 1990s, partly as a result of Paul Gilster’s book of that title (1998), and was firmly aligned to the first phase of the World Wide Web, which was characterised by self-publishing, non-linear reading, the hypertext and searchable content. It has partly been shaped by digital culture and technology (blogging, multimedia, coding, hacking or wikis and social knowledge production), but the plethora of sub-categories (code literacy, data literacy, game literacy) or overlapping concepts (multimedia literacy, information literacy) point to the numerous disciplines (including library and information science, new media and the digital humanities) which have influenced its development.

A substantial body of research about ‘digital literacies’ also relates to school age education and is influenced by strategic debates over pedagogy and between the competing dynamics of centrally mandated policy or practice on the ground/from the ground up (Marsh 2007). The discussion around heavily metricised ‘performance’ versus ‘competence’ models of learning or between rote/scientised and ‘autonomous’ modes of learning is highly relevant here, and Marsh is not alone in criticising what she sees as the superficial ‘appropriation’ of pedagogical material from new literacy practices by official agents, with limited attention to curriculum design or to the media-rich materials now available to us, which ‘dissipate their potential for innovation and transformation’ (Marsh 2007: 267). These disagreements mirror wider differences of opinion about the role of digital culture and technology in research: for some, its role is principally to facilitate (greater access to content or dissemination), whereas for others, it constitutes a paradigm-shifting transformation in how we carry out research and share it with others (Weller 2011). Students and scholars in the humanities already enjoy the benefits of online access and dissemination, although even that stretches current digital-critical competency levels in the humanities, and any further engagement with the digital typically requires a set of critical and technical skills which ask difficult questions of our current analytical toolkit.

While consensual definitions of the term prove elusive, there is an enduring sense that digital literacies are a pressing matter for many at all levels of education. In 2003, Lankshear and Knobel expressed the concern that ‘the future is already here, that we are literally staring at it, but that within school education we are not really engaging with it’ (Lankshear and Knobel 2003: 19). In some countries there has been political support for greater attention to digital literacies – the journal *Wired* cites a UK House of Lords report in 2015 arguing that ‘[t]he incoming government should make sure digital literacy has as important a role as English and maths in the school curriculum’ (Collins 2015), although this is often voiced (or at least received) primarily in terms of vocational or technical skills.

The UK Literacy Association gave digital literacy a different focus, suggesting in 2011 that ‘there is an urgent need to help children develop critical digital literacy skills’ at primary level (Stone 2011: 5). More broadly, Marsh argues that ‘it is essential that schools offer opportunities for all children to become competent and effective analysers and producers of a range of multimodal texts and artefacts’ (Marsh 2007: 268). Likewise, in their foreword to the book *Digital Skills: Unlocking the Information Society* (van Deursen and van Dijk 2014), Michael Thomas, James Gee and John Palfrey voice the common claim that current education practices do not engage ‘learners or teachers with the skills required for the Twenty-first century’, citing the perceived need for greater attention to digitally mediated ‘active’, ‘creative’ and ‘problem-solving’ aspects which align more closely with the knowledge economy of today in “authentic environments” (Thomas et al. 2014: ix).

In a higher education context, Pannapacker argued that ‘digital literacy is going to be as essential as information literacy and critical thinking’ in a blog post for the *Chronicle of Higher Education* in 2009 (Pannapacker 2009),⁷ while Cathy Davidson and Mark Surman in their examination of ‘web literacy’ contend that the open architecture of the Internet sets up a “major educational paradigm shift”, bringing with it a whole set of new issues around collaboration, privacy, personal identity and the culture of interaction (Davidson and Surman 2012). All of this brings particular challenges for the humanities, which need to adapt their methodologies to facilitate knowledge production around complex cultural problems, with attention to a greater diversity of content formats now including databases, text corpora, multimedia and machine code (Suárez 2013). I will now trace some of the main conditions influencing digital literacies in the humanities.

Critical issues and topics

Condition 1: ‘The digital’ in culture and society

I will start my review of the conditions for digital literacies in the humanities by considering some of the main general characteristics of digital culture and their impact on culture and communication in society as a whole. I shall highlight four main areas, discussed next.

Content

Digital culture introduces new dynamics around *content*, which is mutable and dynamic, but also increasingly unstable in comparison to print content. Digital content is easy to produce and easily *reproducible* (and therefore *transferable*). This has economic/structural implications, as content moves from an economics of scarcity to an *economics of abundance* (exemplified by the Freemium cost model) and expectations change regarding *access*, *validation* and *peer review*. We are increasingly used to a number of different devices, platforms

and channels which provide multiple *modes of access* and numerous *affordances*. We are also used to working with a much wider range of content types, which now include online presentations/slideshows, blogs, podcasts, vlogging/videocasts and tweets.

Practices

Practices around digitally mediated communication and knowledge creation are more diverse than they used to be, and associated literacy activities cover a vast range of possibilities, including blogging, editing/uploading images or multimedia, taking part in online debates, creating/consuming digital mashups or filtering/analysing/visualising data. As a result, they are *less bound by disciplinary or professional norms*: this facilitates creative application and connectivity across domains, but it also requires greater agency and reflection by the user in making sense of outcomes. Practice and learning are more closely bound than before, through self-guided learning, practice-based education and project-focused pedagogies.

These practices involve a different set of skills based on participatory culture: Jenkins et al. list skills such as play, performance, simulation, appropriation, multitasking, distributed cognition, collective intelligence, judgement, transmedia navigation, networking and negotiation. Knowledge creation is increasingly *collaborative* and *interactive*, with shared authorship the norm in many settings. Everyone becomes a publisher or a *prosumer* as the divisions between consumer and producer weaken (Jenkins et al. 2009).

The language of digital culture frequently emphasises ‘openness’ and the ‘democratic’ nature of digitally mediated communication, but it is also subject to both new and familiar geographically, culturally and linguistic inflected *dynamics of exclusion and marginalisation*.

Beyond text

Information is *less text-centric*; it is propagated and visualised in a number of different ways, including image-based content, spatial representations and multimedia, which require a new set of critical competencies, including visual, spatial and media literacies. Human communication has always occurred in multiple modes, through spoken or ‘written’ culture – text, images, sound, animation, video, maps – in combination, but digital culture has made the interaction between different forms more dynamic and has given greater emphasis to visual culture, a fact, arguably, which is still not sufficiently reflected in our curricula. Whereas some visual literacy is learned automatically, higher-order visual literary skills require specific attention (Ausburn and Ausburn 1978: 288), and critical awareness regarding the audience, intentionality, authorship/reception, context, production techniques and affective impact of imagery or animated content is crucial to interpret the socially constructed and ideological powers of dynamic and aesthetically appealing content (Bamford 2003), which sometimes feel more ‘immediate’ or ‘direct’.

Humans now need the skills to interpret what is happening on both paper and screen, the latter within a wide range of digital devices, each presenting its own system of gestures and metaphors.⁸ The increasing influence of multimodal texts facilitates interaction and involvement of audiences, creating stronger affective experiences and employing ‘visual arguments’ which work to different logics to narrative-based argument (Jones and Hafner 2012). Visual literacy ‘emerges from seeing and integrating sensory experiences’, involves interpreting ‘ – sometimes simultaneously – visible actions and symbols’ and requires design-based sensitivities (Jones-Kavalier and Flannigan 2006: 9). Some argue that ‘the public appear to

be far more easily convinced by visualizations than by reading text. The implication is that their level of visualization literacy is not as discerning' (Champion 2017: i27, referencing Pandey et al. 2014). Visualisation, then, is a special case often connected to the concept of 'data'. Visualisations have become increasingly popular as forms of argument, but are often under-theorised, methodologically suspect and interpretatively poor (Drucker 2011).

Taken as a whole (and including text-based interpretations of literacy), these new literacies correspond to the modern-day need to produce and interpret, and therefore to code and decode, a varied range of signs, symbols and signifiers, frequently interacting in unpredictable combinations.

Identities

Digitally mediated or online identities are complex and resist simple categorisation. Early categorisations, such as Prensky's 'digital natives' and 'digital immigrants' (Prensky 2001), have long been contested – White and Le Cornu, for example, propose alternative groupings of 'digital residents' and 'digital visitors' to describe online engagement and argue that the terms should not be seen as fixed labels, but rather dynamic identities which we all take on at different stages in our engagement with digital media (White and Le Cornu 2011). Contemporary research explores 'ways of being' which contemplate embodiment, attitudes, gesture and acts (or performance) – with the implication that digital identities consist of fragmented and unstable representations, with an increasingly complex relationship between our offline and online identities.

A strong part of our mediated identities consists in our ability to make connections, and the Internet has developed a strong culture of self-promotion and self-referentiality, in some cultures at least, with social identity performance and impression management (Jones and Hafner 2012: 152) having some influence on our professional identities too, through work-facing social media or impact metrics. The ability to create and maintain virtual profiles and to manage personal identities and data online is increasingly important, and this is creating new economies of participation with their own dynamics of social differentiation. The new 'online affinity spaces' (Gee 2004) created by different social groups are to some extent replicated in professional or learning environments, where people learn new norms of interaction and cultural codes. We should consider, too, the learning experiences in these spaces – how people hear about them, enter as novices, 'serve their apprenticeships' and eventually become experts – which influence learning expectations elsewhere and set up new possibilities for pedagogical engagement. And finally, we need to consider the Anglophone origin of the most commonly used digital tools and platforms, the global role of English language in many countries and the influence that has on communication.

Condition 2: digital scholarship

Digital scholarship functions in this digital cultural space, and so is responsive to the dynamics listed earlier, but it deserves some dedicated attention because it shapes *what* is taught and *how* it is taught in academic institutional settings, and while academia increasingly competes with other actors and processes in its knowledge legitimisation claims, it still represents an important and discrete site for knowledge production with its own norms and architectures of participation.

There are many reasons why scholarly practice has been more cautious than other areas of human professional life to adapt to digital culture, methods and tools. In *The Digital*

Scholar, How Technology Is Transforming Practice (Weller 2011: 41–43), Weller explores the many ways in which scholarship is susceptible to the impact of digital culture and argues for us to draw on lessons from other professional sectors (such as the media or music industries) and to pay careful consideration to the impact of these changes on the emerging media ecologies. In examining the nature of scholarship, he draws on Unsworth’s scholarly primitives (discovering, annotating, comparing, referring, sampling, illustrating, representing) (Unsworth 2000), the task-based categorisation of Palmer et al. (searching, collecting, reading, writing and collaborating) (Palmer et al. 2010) and Boyer’s definition of scholarship (discovery, integration, application, teaching) (Boyer 1990) to examine changing practices, and in so doing, he explores a whole range of potential implications for modern-day literacies, including the digitally mediated ‘networked research cycle’, ‘permeable boundaries’, openness in education, dissemination and their impact on scholarly reward systems. It would seem sensible to draw on these more closely when designing digital literacy provision in higher education, or at least to be explicit about the scholarly ‘primitives’ which are being engaged each time we do so.

We do not have space to explore these issues in depth here, except to note that digital scholarship currently operates at the boundary between two sets of competing dynamics: namely principles such as careful reflection, expert peer review and stable publication/citation, on the one hand, and factors such as open/social knowledge creation, post-publication review and dynamic dissemination, on the other. Scholarly publications are still largely produced according to print era workflows and economics, although new forms of creation, consumption and reception are increasingly evident, and ‘social’ online communication introduces new dynamics of interaction which are not currently formalised within the structures of digital scholarship.

So digital scholarship has to contend with a series of contradictory currents (immediacy, public impact and the collaborative reflex versus rigour, stability and the favouring of single-author publications) which profoundly influence the conception and delivery of digital literacies in a university setting. Some of the low-hanging fruit of digital culture has been largely absorbed into institutional structures (electronic communication, online discovery, digital reading lists), but other aspects remain very much open for discussion, particularly in the humanities, such as perceptions about the relationship between the humanities and the tools it uses, or the (sometimes heated) debates about open access in scholarly publishing. This is important, because there is a big difference between teaching English using social media or other corporate-born proprietary tools and teaching it using open-source and community-driven tools, where the design and construction of the tools are part of the curriculum. We now move to approaching these issues from a more specifically humanities perspective.

Condition 3: digital (literacies for the) humanities

New technology, social networks and research in humanities

Some argue that the humanities should engage more closely with ‘social’ networks and technologies, within a vision of universities which re-affirms their role as ‘social agents’ and ‘generators of cultural production’ (Baraibar and Cohen 2012: 157, *my translation*). Baraibar and Cohen argue that knowledge generated within the humanities has the potential to be re-applied or re-interpreted in a way which is not as relevant in the sciences, where each stage of knowledge acquisition often becomes obsolete more quickly.

In this interpretation, where research can become part of a wider conversation, involving wider publics, and with different dynamics of information flow, digital literacies would need to not only account for studying online databases and archives but also for digitally mediated research as a process (or as ‘a conversation’, in the words of the authors), including collaboration, the opportunities presented by the ‘public’ humanities or the dynamics of digital knowledge production.

Changes to knowledge in the classroom

It is tempting to think of the problem in terms of ‘digital tools’ that can be learned as an additional element which does not rupture the basic fabric of humanities teaching/learning, but new digital epistemologies challenge us to rethink the nature of knowledge in the humanities classroom and to reassess ‘the world (objects, phenomena) to be known’, our ‘processes of ‘coming to know’, ‘the constitution of ‘knowers’ and “changes in the relative significance of, and balance among, different forms and modes of knowledge’ (Lankshear and Knobel 2003). This points to a more radical view of digital literacies, which engages not just with the raw materials of scholarship, but also with its architecture, flows and mind-sets.

Introducing the digital humanities

Enter the digital humanities, a transdisciplinary field which promotes a critical and ‘hands-on’ engagement with digital methods, tools and infrastructure, and which often encourages reflection not just on *what*, but also on *how* and *where* digital literacies might be developed. While by no means the only site for experimentation and critical reflection about the pedagogical applications of digital culture and technology, the digital humanities offer some useful vantage points from which to view digital literacies in the humanities, and in English studies in particular.

‘What is Digital Humanities and What’s It Doing in English Departments?’ asks Kirschenbaum in the title of his influential essay about the field (Kirschenbaum 2010: 3), who goes on to argue that the scholarly and professional apparatus it has developed has closer alignment to English than to any other discipline. This has been contested by some as representing a North American perspective (Dacos 2013), but it is certainly true that digital humanities is particularly strong in English departments in the United States, as witnessed by historic projects like *Women Writers Online*, the *Rossetti Archive* and *The Orlando Project*. Kirschenbaum argues that the field’s work in various areas – including textual scholarship, linguistics, stylistics, electronic literature, composition, archives, e-reading and data mining – make it a natural match to English as a subject and that its encouragement to make scholarship ‘publicly visible’ is something any English department would wish to embrace (Kirschenbaum 2010: 9). As we shall see, English studies have thus shaped many of the ‘DH’ curricula which have been developed over the years, and this has had an important effect on how digital literacies are understood by the field.

The fact that the digital humanities resist stable definition has both benefits and disadvantages. The fluidity and dynamism of the term allows DH to react quickly to changes in the digital cultural landscape (clearly an advantage in the sense that learners need to adapt quickly to new tools and new paradigms), but it also makes it challenging to reliably define its limits and scope, which makes it challenging to substantiate claims about, or criticisms levelled, at the field. Some, such as Antonijević, acknowledge a difference between the digital humanities as a field and the broader concept of digital scholarship in the humanities and

argue for the benefits of having clearer boundaries (Antonijević 2015). Her in-depth ethnographic study of digital practices among a wide range of humanities scholars highlighted a number of challenges in the disciplinary transformations which result, including *research capacity* (the condition of methods and tools and our relationship with them), *knowledge capacity* (knowledge exchange, ethos of collaboration, openness to innovation), *value capacity* (scholarly attitudes towards ‘the digital’, uncertain reception towards ‘collaborative’ or ‘digital’ outputs, review and validation mechanisms) and *technical capacity* (useful features, degree of humanities influence over technical design) (Antonijević 2015: 73–102).

In ‘DH as a Humanities project’, Svensson (2012: 43) asks why the digital humanities might be a ‘legitimate place’ to think about the humanities. Arguing that ‘DH’ is a ‘useful one-stop manifestation’ of the humanities as a whole, which at times seems to question traditional structures in the humanities, Svensson (2012: 54) proposes that the digital humanities is best understood as a ‘meeting place, innovation hub and trading zone’. Together with Theo Goldberg, he later expands on this point to suggest that it offers the humanities an opportunity to engage more closely with technical disciplines such as engineering and computing science; to realise ‘productive’ interactions with wider publics; to recalibrate knowledge production in the humanities to give greater emphasis, for example, to ‘multimodal expressions’; and to reassess the role of the humanities in rapidly changing times (Svensson and Goldberg 2015: 5–6). This appeal to rethink the formal structures of the humanities, or universities more generally, is not uncommon in this field: Cunningham et al., for example,

want to encourage readers to re-imagine the way higher education works: from the way it is supported, to its disciplinary divisions, to the possibilities and limitations inherent in its pedagogical mission, to the need to innovate in order to answer those needs in the twenty-first century.

(Cunningham et al. 2008)

The networked alliance HASTAC puts many of these ideas into practice, aiming to change ‘the way we teach and learn’ through shared ‘news, tools, research, insights, pedagogy, methods, and projects’.⁹

We might, similarly, ask why the field of digital humanities is a legitimate place to think about digital literacy in humanities. There is little evidence to suggest that digital humanities have engaged with the concept of (digital) ‘literacies’ in a systematic way so far. Searching two of the main journals, *Digital Scholarship in the Humanities*¹⁰ and *Digital Humanities Quarterly*,¹¹ for the term (in singular or plural forms) one finds relatively little evidence of engagement, and the usage which exists covers a variety of terms (including terms already covered here, in addition to the term ‘sonic literacy’ (Barber 2016), without favouring any in particular.

It might be argued that DH is suspicious of the term because it wants to avoid accusations that it simply engages in ‘skills training’, but equally, an argument could be made that DH is already promoting digital literacies at the more ‘critical’ and practical end of the spectrum, and so engaging more fully with the discourse of digital literacies would serve to render the field’s strategic educational role more explicit. In any case, DH is well placed to contribute to this agenda for a number of reasons, including the fact that it already takes a mediating and transdisciplinary role in research; it can operate at a number of different locations within institutions; and it represents an ongoing, dynamic and critical debate about the relationship of the humanities to rapid evolutions in digital culture.

Condition 4: learning environments

The final condition for digital literacies relates to learning environments. Many aspects of our current literacy needs are not satisfactorily attended to by our current curricula, which are still largely built on print culture principles: it is clear that teachers, learners and pedagogy are still very much in transition. New media bring new frameworks for learning, with new possibilities for reading, assessment and tracking but also new risks (Stone 2011).

There is also substantial evidence that our economies of attention are changing (Lankshear and Knobel 2003), a fact not sufficiently represented in pedagogical strategies at present, and not enough consideration is given to the *science* of attention, where authors such as Davidson use advances in cognitive science to propose a re-organisation of twentieth-century educational institutions and course structures. In focusing on new forms of learning through gaming, self-evaluation and collaborative activities, she argues, we are better placed to engage with the modern learner by harnessing natural behaviours and interactions (Davidson 2012).

Current contributions and research

Components of digital literacy

So far we have seen many ideas about what we might include as components of digital literacies, but how can these be consolidated into a manageable list? There is a wide spectrum in how people approach digital literacies in the humanities, which ranges from ‘basic’ everyday skills, applicable to everyone, to more specialised skills, dictated by someone’s domain of study (e.g. distant reading of nineteenth-century literature or geospatial models for studying language development).

General digital literacies

At one end of the scale, digital literacies in the humanities may draw on broader competencies and skills required in today’s learning environments. Jisc highlight the relationship of ‘information’, ‘media’ and ‘data’ literacies within wider ‘digital capabilities’ which include functional skills, critical use (literacies), creative production, participation, development and self-actualising.¹² Some of the competencies mentioned by Jones and Hafner include:

- Search and discovery across large amounts of information. This also implies working with search engines, understanding the role of technical, social and personal filters and algorithms in shaping our digital engagement.
 - Navigating pathways through complex collections of connected content.
 - Connections across concepts and fields of expertise.
 - The production and interpretation of audiovisual or multimedia material.
 - The ability to construct identities, interact and collaborate in online social communities.
- (Jones and Hafner 2012: 1, 28–30)

There is a strong sense of remixing, curating which traverses these themes, and there is a generative, creative element which perhaps deserves particular attention. It is challenging to settle on a stable list of digital competencies which works across disciplines, academic levels and educational cultures, but one of the most convincing is provided

by Hafner et al. (Hafner et al. 2015), who summarise earlier work by (Jones and Hafner 2012) in a list of five pairings by (1) dimension and (2) focus:

- Doing/actions in the physical world
- Meaning/forms of representation
- Relating/patterns of interaction
- Thinking/experiencing and thinking about reality
- Being/social identity
- Focus on tools

At the other end of the scale are the more hands-on approaches to the building blocks of digital culture, an engagement with digital methods, tools and coding which largely occurs under the umbrella of DH. Tools have been a common starting point for this engagement, because they allow for a relatively accessible path towards understanding some of the inner workings of digital culture without the need to learn code from the start, and so they have been a crucial part of digital literacies as these might be understood in a DH context. Digital humanists have created a huge number of tools,¹³ covering most areas of the humanities, and with a wide array of associated research techniques, objects and activities. This body of practical work, while sometimes criticised for its lack of user engagement (Warwick et al. 2008), constitutes an important source for studying practice-based digital literacies for a number of reasons: it is produced by a cohesive community of practice (Lave and Wenger 1991) under the DH umbrella; it bridges the divide between the humanities and a sociotechnical infrastructure, which is otherwise largely resistant to cultural criticism; and it encourages the kind of pedagogical experience (learning by doing, experiential learning) which are becoming key features of the emerging educational landscape. It also forces us to contemplate *method* explicitly, the second key element in a DH view of new literacies.

Main research methods

Two quite different approaches to digital methods are evident, one of which concentrates on digitised surrogates of analogue content (the digitised past) and the other which focuses on born digital content (the digital present).

Digital methods for digitised surrogates of analogue content

The first of these approaches is firmly rooted in the digitisation of pre-digital content which took off in the 1990s – this is also closest to traditional DH research trajectories, in Europe at least, and overlaps with other common themes such as tool creation, digital editing and building collections/corpora. This has often been closely connected with cultural heritage institutions and projects and has incorporated curatorial and archival practices of curation, preservation and digital asset management (DAM). Over time, this research landscape has often evolved into e-research agendas, exploring connective data, geospatial and big or linked data perspectives on research objects whose methods sit at the intersection between humanities disciplines, collections and researchers, bringing together digital content, tools and methods (Hughes et al. 2015: 150–152). One attempt at capturing a taxonomy of research activities is provided by the TaDiRAH resource,¹⁴ which includes terms such as ‘capture’, ‘enrichment’ and ‘storage’ – the kinds of digital literacies favoured in this approach clearly

emphasise the bridge between the humanities and library and information science agendas, in scholarly terms, and the galleries, libraries, archives and museums (GLAM) sector in the professional sphere. The kind of ‘Methods in Practice’ (Hughes et al. 2015: 157) promoted by Network for Digital Methods in the Arts and Humanities (NeDiMAH) give some hint of the common points of emphasis in European DH – ‘Space and Time’, ‘Information Visualisation’, ‘Linked Data and Ontological Methods’, ‘Building and Developing Collections of Digital Data for Research’, ‘Using Large-Scale Text Collections for Research’ and ‘Creating Digital Editions’.¹⁵ The relationship between these research methods and digital literacies has not been sufficiently studied up to now, but it is an area which would benefit from greater collaboration between the higher education and GLAM sectors.

Digital methods and born digital content

Another, very different, approach to digital methods concentrates on born digital content, in particular in relation to the Web. Following on from his earlier work about ‘Preferred Placement’, which explored the politics of the Web and their impact on knowledge production (Rogers 2000: introduction), Rogers and others have charted a new course for digital methods, which now include studying the filters, flows and reputational indicators of the Web through studies of link politics, Twitter interactions, Wikipedia edits and search dynamics (Rogers 2013). Digital methods, in this definition, serve ‘as a counterpoint to virtual methods, which typically digitise existing methods and port them onto the Web. Digital methods, contrariwise, seek to learn from the methods built into the dominant devices online, and repurpose them for social and cultural research’.¹⁶

What applications does this have to humanities disciplines such as English studies, history or comparative literature? At first glance, not a lot for now. On the face of it, digital methods are more easily applicable to disciplines with a strong presence of post-Internet objects of study, either because the subject matter is born digital, or more commonly, because its process of reception, analysis and further dissemination are digitally mediated. On closer examination, however, this does provide the humanities with some interesting challenges: How will it transform reception studies, for example? How will we trace the transmission history of a text once it starts to enter (and leave) digital realms? Will the heavy emphasis on ‘data’ alter the currently distant (and often outright suspicious) attitude in many areas of the humanities to ‘data’? Internet study methods such as sampling, Internet ethics and digitally mediated research design (Hewson et al. 2016) will no doubt change how we construct research programmes at a higher level, but will they have a stronger bearing on the substance of humanities research itself? We can already see a few examples of its influence on discussion of humanities methods,¹⁷ but in most cases, at present ‘digital’ and ‘non-digital’ methods seem to operate in parallel, with little work to formalise their interactions from a theoretical point of view, even in studies of contemporary cultural and historical subjects. A ‘digital literacies’ agenda needs to address this point.

Code, coding

Code literacy is sometimes raised as an important part of a digitally aware education. In ‘Program or Be Programmed’, Rushkoff encourages us to code so that we can ‘gain access to the control panel of civilization’ and to find our place ‘in the new cybernetic order’ (Rushkoff 2011: 7–8). Rushkoff claims that over time, as interfaces have simplified our interaction with digital devices, we have actually become further distanced from their underlying code

and architecture, thus creating ‘black boxes’ which are hard to analyse/comprehend, thus rendering us vulnerable to manipulation by commercial interests.

There have sometimes been heated debates about whether or not ‘coding’ (Ramsay 2011c) (or more generally ‘building’) constitutes a significant element in DH (Ramsay 2011a). Formal DH sometimes include programming as a core element, but where they do, this is almost universally from beginner level and is aimed at humanities researchers and students. It is also important to signal that the focus of such learning is very different from computing science courses covering the same topic, since the outcomes are very different – the standard expectation is not that a DH coding course will normally lead to a programming role, but that it will lead to complementary analytical or managerial roles. This adds a professional perspective to Rushkoff’s comments – namely that some understanding of code not only has social implications but also has career benefits.

Nick Montfort makes a different point about the distinguishing features of coding in the humanities. Directing criticism at DH volumes such as *Digital Humanities* and *Debates in the Digital Humanities* for their non-inclusion of discussion about coding (Montfort 2016: 98), he argues for a kind of ‘exploratory programming’ which is quite distinct from the focus of other courses which are shaped around software delivery. This view on coding takes ‘computation as a way of inquiring about and constructively thinking about important issues’ and argues for ‘programming’s cognitive, cultural and social value’ (Montfort 2016: 98). Rather than seeing the computer as a ‘power tool’, he says we should use it as ‘a sketchpad, sandbox, prototyping kit, telescope and microscope’, a tool for exploration rather than mere implementation (Montfort 2016: 108). In this interpretation, the intellectual agendas of the humanities mould the digital tools and their affordances, rather than the other way around. This follows Kirschenbaum’s argument that computers should be ‘understood as engines for creating powerful and persuasive models of the world around us’ and that we need to have some understanding (from a humanities perspective) of their rules, models and algorithms in order to harness their discursive power adequately (Kirschenbaum 2009).

This talk of ‘computational literacy’, or ‘iteracy’ as Berry calls it (Berry 2012: 15), has a long history, going back to the 1960s, when some argued that all undergraduate students should be taught programming (Vee 2013: 43). Referencing testimonies on code.org,¹⁸ a resource supported by various high-profile digital entrepreneurs, Vee finds that some describe code as an ‘issue of civil rights’, but she is critical of the lack of social criticism in relation to this use of the word ‘literacy’, warning against the unthinking adoption of computing science-based values and advising us to pay attention to gender and other imbalances (Vee 2013: 43–44, 58). This is emphatically coding humanities style, but any attempt to present an alternative to the view of code literacy as something fundamentally skills-based, belonging to the ‘technical sphere’ and culturally neutral, faces numerous challenges of image and expectation, not least from the humanities themselves. And yet code literacy, many in the DH would argue, is important. In case this seems like an impossible task, it is worth pointing out that we have been teaching coding as a central (and compulsory) part of our MA in DH for over ten years to people who mostly came from a humanities background and the vast majority of whom had no previous experience,¹⁹ which will lead to the outcome that they will ‘[b]e able to scope, build and critique practical experiments in digital scholarship with an arts/humanities focus’.²⁰

Zundert et al. argue that ‘code’ and ‘text’ need to be seen as ‘variations within a continuum of literacies’ (Zundert and Haentjens Dekker 2017: i2), while a proper understanding of code helps us to understand how tools work, to explore the ‘computational acts’ which can open up new forms of interpretation, argues Ramsay (Ramsay 2011b).

What does this mean for the humanities and some of its central scholarly frameworks, such as the hermeneutic method? In ‘Hermeneutica: Computer-Assisted Interpretation in the Humanities’, Rockwell and Sinclair present us with a theoretical and methodological bridge ‘across the divide between writing code and writing text’ (Rockwell and Sinclair 2016: 5), in what they call the ‘Agile Hermeneutics’ collaborative practice. This process of thinking through a problem using digital tools and technology has important implications for digital literacies.

Critical literacy

As before, the ‘critical’ nature of digital literacies is crucial here, and Jones and Hafner warn us to be wary of the technodeterminism of ‘digital sceptics’ and ‘digital utopians’ alike for their lack of attention to the ideological element in the design of technology: the ‘social contexts in which it is used and the intentions of those who use it’. In doing so, they argue, we need to be aware of affordances and constraints, social conventions which develop in usage, the differences between open and closed systems, hidden forms of bias (for example as a result of quantification), the unequal power relations which govern technology use and the ease of use/ensuing ‘investments’ we make in particular tools and technologies (Jones and Hafner 2012: 99–100).

This seems like fertile terrain for the humanities: many have pursued similar arguments in relation to the DH and digital scholarship in the humanities more generally: ‘The humanities, with their broad intellectual perspective and critical sensibility, can help us beyond the latest fads and think more deeply about the role of technology in the modern world’ (Haigh 2014: 28). So, while we would do well to reconsider humanities pedagogy in terms of digital literacies, we would also benefit from applying a critical humanities perspective to digital literacies: to think how and why they evolve and what action we want to take about that.

How to acquire digital literacies

I have demonstrated the range of possible components which could be considered important building blocks for digital literacies in the humanities. The question of how they are acquired presents an equally diverse range of responses. Is digital literacy an academic programme, a learning module, a skills course or simply part of everyday learning and development? There is evidence for all of these options, and more.

Self-taught digital literacies, learning by doing

While it is extremely hard to gauge the shape and dynamics of self-learning in any measurable way, there is no doubt that many learn digital competencies on their own and *by doing*. This ‘just-in-time’ form of learning has the advantage that it responds to real needs (rather than abstract or generalist concerns), is often carried out in context and sometimes involves friends and colleagues in communities of practice. Ethnographic research by Antonijević into the ‘knowledge capacity’ of humanities scholars certainly seems to confirm this individual approach, but her study also suggests that this ‘organic learning’ is a rather isolated affair, with little support for formal instruction and concerns that ‘digital skills’ are undervalued (Antonijević 2015: 85). While there are some humanities-specific resources such as the Programming Historian,²¹ most technical resources which people use (such as code tutorials for Hypertext Markup Language [HTML], JavaScript or

Python) are produced to a commercial agenda, with little consideration given to humanities concerns such as representation, interpretation or context.

Workshops

One response to the limited learning opportunities within the formal curriculum (or staff development programmes) is to run practical workshops, and this is an area where the DH has been particularly active, with one-off events, hackathons, ‘moots’,²² pre-conference workshops²³ and summer schools or institutes.²⁴ These events are often presented in friendly, accessible terms (‘the DHSI prides itself on its friendly, informal, and collegial atmosphere’)²⁵ and aimed at a non-technical audience. The advantage of these is that they give quick practical introductions to particular tools and methods, but there is a risk that little time is given to a process of reflection, because the attention of the learner is focused on the technical process of the tool at hand. It is quite common, moreover, for learners to be initially captivated by the initial ‘thrill’ of new tools, but that crucial next step where they are challenged to consider their affordances and limitations in relation to existing scholarship, and to articulate their role within a wider research ecosystem, often happens outside of the learning environment, with diminished opportunities for connection to theory or reflective practice.

Formal courses

It could be argued that specialist ‘digital’ programmes (such as the BA in digital culture or MA in digital humanities which we teach at King’s College London) offer a solution to this, bringing ‘digital theory and practice to the study of human culture: from history, English and music to museums, digital publishing and beyond’²⁶ in formal academic programmes designed to evaluate new visual, spatial and textual literacies. The appeal of these programmes is often in their vocational inflection, fusing academic perspectives with project-based work which sometimes allows for collaboration, problem-based learning and experience working with partners outside of academia.

These programmes are popular with learners wishing to choose ‘digital’ perspectives as a primary area of study, but they do not deal with the broader challenge of integrating digital literacies into the humanities curriculum. As I observed earlier, it is hard to gauge the extent to which digital literacies figure within formal learning environments in the humanities, but there is little evidence at present that they form part of any coherent plan, and where terms like ‘digital literacy’ appear in strategic education documents they are usually left undefined; as malleable signifiers which can be reshaped according to immediate context, rather than being given strategic agency.

Literacy frameworks

What kind of framework would we need to implement a strategy of digital literacies in the humanities? Summey’s analysis of frameworks for developing digital literacies, while aimed at a different audience, provides some useful elements which are likely to form a part of our response to this, such as room for experimentation, the value of collective intelligence, peer mentoring and personal learning networks (Summey 2013). Do we need humanities equivalents of standards such as the ISTE – National Educational Technology Standards²⁷ or the American Association of School Librarians – standards for the

twenty-first-century learner²⁸ mentioned by Summey (Summey 2013) in his analysis? And what are the wider pedagogical implications?

Wider pedagogical issues

It is beyond the scope of this chapter to examine the potential impact of digital literacies on the learning process in depth here, but it is important to observe that they will both influence and be influenced by the changing pedagogical landscape, with increased attention to self-study, collective knowledge, peer support/mentoring and the ‘flipped classroom’. Some argue that there is a disjoint between how teachers and students experience digital culture: ‘[a] common scenario today is a classroom filled with digitally literate students being led by linear-thinking, technologically stymied instructors’ (Jones-Kavalier and Flannigan 2006: 8). The reality is probably more complex, with differentiation as acute within both groups as between them, but there is no doubt that the roles of ‘teacher’ and ‘learner’ are being challenged and will continue to evolve in response to new expectations and experiences around media, content and interaction.

Gillen points out that ‘digital learning practices are intrinsically dynamic’ and argues for particular attention to the dispositions of learners in light of this: their ability to adapt quickly and continuously to new circumstances, to maintain critical perspective and to develop ‘resilience’ in overcoming obstacles and to navigate a heterogeneous array of cultural practices (Gillen 2014: 154).

Many digital courses have, from early on, involved project work (Jessop 2005: 296), and while this risks undermining the deep individual reflection which is a crucial element in many areas of humanities research, it finds echoes in some aspects of the *pedagogies of abundance* outlined by Weller, such as resource-based learning (RBL) and problem-based learning (PBL) (Weller 2011: 91).

In the DH, the importance of enabling students to learn to become *builders* within projects, even at undergraduate level, is a common feature, and Mauro goes as far as to argue that these projects ‘must emphasize iterability, openness and extensibility’ (Mauro 2016: 379) which can lead to portfolios of potential value in their professional lives later.

This mirrors an important shift from content- and process-based modelling to problem-based modelling identified by Siemens and Sayers (Siemens and Sayers 2015) and represents the kind of ‘purposefully decentered learning’ which has the potential to transform how learners see themselves (Lankshear and Knobel 2003). This pedagogical shift to some extent follows the model of informal and interactive learning, which Gee evokes in his definition of *affinity spaces* (Gee 2004), encouraging students to find stronger connections between their in-class and out-of-class experiences and learning processes (Hafner et al. 2015).

Cathy Davidson describes a hypothetical final exam where students are required to create a collaborative online project involving learners, teachers, other researchers and the wider public to solve a humanities research problem. Her argument is that most people simply do not have the requisite skills to take on many complex cultural challenges of today, and so we need to ‘collaborate by difference’, bringing together different skillsets and intellectual cultures in order to overcome major intellectual and practical challenges together (Davidson 2015: 131, 136–137). This is the other key pedagogical challenge – collaborative learning – which, on the one hand, presents numerous challenges in terms of institutional structures (for example, around assessment) but which also provides many opportunities to re-think learning with digital literacies in mind.

Future directions

I have not aimed to provide an exhaustive list of ‘digital literacies’ here, instead preferring to focus on those themes which are well developed or might form part of a broader and cohesive programme of literacies appropriate to learning in the humanities at this moment in time. A case could be made here for including network literacy or for including gaming literacy for its adaptability to level, self-monitoring, promotion of engagement, instant feedback and role-playing benefits (Davidson 2012: 144–148), although its benefits are less clear at higher education than at school level. Perhaps a clearer case can be made for data literacy, broadly characterised by Frank et al. as the ‘ability to understand and use data’ and which includes processes of discovery, selection, evaluation, validation, manipulation and visualisation (Frank et al. 2016). While accepting that multiple definitions of data literacy may be appropriate, they suggest that each context may require its own definition, and one such context might be the humanities, where more critical perspectives seem urgent (and not just in relation to ‘big data’ (boyd and Crawford 2012)): it is still far from clear how data operate as a research object more generally in the humanities (Borgman 2015).

Digital literacy and digital inclusion

Frank et al. raise ‘the exciting possibility that data literacy will be the first “literacy” whose dominant model comes from the global south’ (Frank et al. 2016: 7). This seems rather optimistic in current conditions, and indeed some argue that digital divides are growing in many aspects. Van Deursen and van Dijk argue that digital divides operate across four stages: motivation, material access, skills access and usage (van Deursen and van Dijk 2014: 2), and digital inclusion is a function of broader dynamics between digital university, digital world and digital user (Thompson et al. 2014), but the particular dynamics of global digital literacies in the humanities merit further study and attention, including understanding better distinctions between ‘surface’ and ‘deep’ webs, the uneven impact of digital devices and personal/cultural relationships to digitally mediated knowledge (Sanz and Goicoechea 2012: 332).

Conclusion: digital literacy as a programme in the humanities?

How might we bring all of the digital literacies, and their constitutive elements, together? Should we aim to bring together all of the different content strands into one cohesive humanities programme to ensure that the ‘newness’ of digital literacies is covered properly; should we approach the topic by discipline (digital literacies for English studies, for example); or should programmes evolve organically, as a function of local circumstances? How can we ensure that ‘digital incursions’ into the humanities curriculum are approached in a manner which demonstrates the enduring relevance of non-digital skills and competencies such as critical thinking, rather than leading to them being undermined? And what association should there be with digital literacies at other stages of education? How, for example, can we connect this discussion to debates about digital literacies at primary and secondary level so that we are building on expertise and awareness gained there?

Some contend that talk of ‘digital literacies’ in itself is dangerous, that the ‘differentiation’ this implies is artificial or over-emphasises technical factors over social ones (Gillen 2014) and we certainly always need to be wary of the dangers of technodeterminism, but I would argue that special attention to the ‘digital’ within wider literacies is necessary, for

now at least, due to the lacunae which exist in our current understanding of, and programmes in, digital literacies in the humanities.

The rise of DH centres or units in some countries may help to shape an agenda to address this, and over time, while some areas (such as coding) are likely to need special attention, the growing competence of new cohorts of teachers and learners is likely to raise the technical bar and allow us to focus less on technical issues and more on pedagogical needs. In the short term it is likely that the answer to this will depend largely on local context (expertise and research interests of more technically minded staff and departments), but as digital methods simply become part of the furniture in the humanities, teaching and researching the ‘otherness’ of digital is likely to become less urgent. Students should be as comfortable using digital tools for creative writing, studying postcolonial dynamics in digital environments or interpreting/annotating digital editions of literature from the Romantic period as they are exploring their non-digital equivalents.

The challenge will be how to manage the constant change which ‘digital’ represents – new devices, platforms, tools and content formats – and a crucial element in the new literacies people develop will be how to adapt to new innovations, without sacrificing our critical faculties to the digital sublime (Mosco 2005). I have attempted to show in this chapter that digital literacies represent something much more significant than learning how to use WordPress (or for that matter coding in Extended Markup Language [XML]), as imagined by some, and should be something which the humanities (and the DH) should care about. It is partly about how learners are equipped to assess, and where necessary learn, new tools and paradigms quickly, on the go, with a strong element of self-learning/peer support. But above all it is about how we can make digital methods and tools part of the creative, expressive and literate arsenal of the twenty-first-century learner. In the English-speaking world, it only seems natural that English studies, with its close traditional connections to the concept of literacy through the ages, should play a significant part in this learning transformation.

Notes

- 1 By ‘humanities’ I am referring to subjects such as ancient and modern languages (including English), philosophy and the performing arts. For some useful definitions, see www.4humanities.org/2014/12/what-are-the-humanities/
- 2 www.gov.uk/government/publications/national-curriculum-in-england-english-programmes-of-study/national-curriculum-in-england-english-programmes-of-study National Curriculum/DfEE
- 3 <https://ukla.org/about>
- 4 www.internetworldstats.com/stats.htm
- 5 http://ec.europa.eu/eurostat/statistics-explained/index.php/Digital_economy_and_society_statistics_-_households_and_individuals#Internet_access 85% internet access households in EU-28. The highest proportion (97 percent) of households with Internet access in 2016 was recorded in Luxembourg and in the Netherlands.
- 6 www.digitalnewsreport.org/
- 7 www.hastac.org/blogs/nancyholliman/2009/12/30/mla-and-digital-humanities
- 8 www.scientificamerican.com/article/reading-paper-screens/
- 9 www.hastac.org/about-hastac
- 10 <https://academic.oup.com/dsh>
- 11 www.digitalhumanities.org/dhq/
- 12 www.jisc.ac.uk/rd/projects/building-digital-capability. See also www.jisc.ac.uk/guides/developing-digital-literacies
- 13 See the old Digital Research Tools (DIRT) list, at <https://dirtdirectory.org/>, for example.
- 14 See <http://tadirah.dariah.eu/vocab/index.php> TaDiRAH – Taxonomy of Digital Research Activities in the Humanities.
- 15 www.nedimah.eu/

- 16 <https://wiki.digitalmethods.net/Dmi/SummerSchool2017>
- 17 www.history.ac.uk/projects/digital/born-digital-big-data-and-methods-history-and-humanities
- 18 <https://code.org/>
- 19 www.kcl.ac.uk/study/postgraduate/taught-courses/digital-humanities-ma.aspx
- 20 www.kcl.ac.uk/artshums/depts/ddh/study/pgt/modules-2017-18/core-modules/7AAVINDH.aspx
- 21 Afanador-Llach, Maria José, Antonio Rojas Castro, Adam Crymble, Víctor Gayol, Fred Gibbs, Caleb McDaniel, Ian Milligan, Amanda Visconti, and Jeri Wieringa, eds. *The Programming Historian*. 2nd ed., 2017. <http://programminghistorian.org/>.
- 22 www.ahrc.ac.uk/documents/publications/digital-transformations-moot-19-november-2012/
- 23 <http://dh2016.adho.org/workshops/>
- 24 <http://linhd.es/p/dh-summer-2017-lang-en/>
- 25 <http://dhsi.org/>
- 26 www.kcl.ac.uk/study/postgraduate/taught-courses/digital-humanities-ma.aspx
- 27 www.iste.org/standards/standards
- 28 www.ala.org/aasl/standards/learning

Further reading

- 1 Battershill, C. and Ross, S. (2017). *Using digital humanities in the classroom. A practical introduction for teachers, lecturers and students*. London: Bloomsbury.
This book is a practical manual in how to integrate digital humanities in teaching and learning. Using practical examples, also available on the companion website www.teachdh.com, it covers the whole learning cycle from inception to classroom to assessment and beyond.
- 2 Constance Crompton et al. (eds.). (2016). *Doing digital humanities: Practice, training, research*. New York, NY: Routledge.
This edited volume provides an overview of digital humanities methods organised into four sections: foundations; core concepts and skills; creation, remediation and curation; and administration, dissemination and research.
- 3 Davidson, C. (2017). *The new education*. New York: Basic Books.
This book explores how higher education has been transformed by a series of social upheavals and examines the role of technologies in new forms of learning. It provides practical tips for getting the most out of your course and for making the classroom a place of active student-centred learning.
- 4 Jones, R.H. and Hafner, C.A. (2012). *Understanding digital literacies: A practical introduction*. Milton Park, Abingdon, Oxon; New York: Routledge.
This book introduces the concept of digital literacies, providing historical context, practical examples and critical tools to situate these new literacies in our teaching and learning. It provides useful case studies, discussion activities, further visual/reading resources and a companion website.
- 5 Openmethods: Highlighting digital humanities methods and tools. <https://openmethods.dariah.eu/>
This is an online resource created by the Humanities at Scale project as part of a DARIAH project initiative looking into digital research infrastructure for the arts and humanities. It provides curated content about digital humanities methods and tools with a multilingual perspective.

References

- Antonijević, S. (2015). *Amongst digital humanists – an ethnographic study of digital knowledge production*. Basingstoke, Hampshire and New York: Palgrave Macmillan.
- Ausburn, L. and Ausburn, F. (1978). Visual literacy: Background, theory and practice. *PLET* 15(4): 291–297.
- Bamford, A. (2003). *The visual literacy white paper*. Audio-Visual Learning. Adobe Systems.
- Baraibar, Á. and Cohen, S. (2012). Nuevas tecnologías y redes sociales en la investigación en Humanidades. *La Perinola*: 155–164.
- Barber, J.F. (2016). Sound and digital humanities: Reflecting on a DHSI course. *Digital Humanities Quarterly* 10(1).

- Berry, D.M. (2012). Introduction: Understanding the digital humanities. In D.M. Berry (ed.), *Understanding digital humanities*. Houndmills, Basingstoke, Hampshire and New York: Palgrave Macmillan, pp. 1–20.
- Bolter, J.D. and Grusin, R. (2000). *Remediation: Understanding new media*, 1st ed. Cambridge, MA: The MIT Press.
- Borgman, C.L. (2015). *Big data, little data, no data: scholarship in the networked world*. Cambridge, MA: The MIT Press.
- boyd, d. and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5): 662–679.
- Boyer, E.L. (1990). *Scholarship reconsidered: Priorities of the Professoriate*, 1 ed. San Francisco: Jossey Bass.
- Champion, E.M. (2017). Digital humanities is text heavy, visualization light, and simulation poor. *Digital Scholarship in the Humanities* 32(suppl_1): i25–i32, 1 April.
- Clark, C. and Douglas, J. (2011). *Young people's reading and writing: An in-depth study focusing on enjoyment, behaviour, attitudes and attainment*. London: National Literacy Trust.
- Collins, K. (2015). Digital literacy should be 'as important as English and maths'. *Wired*. Available at: www.wired.co.uk/article/lords-digital-skills-report (Accessed 26 April 2017).
- Cunningham, R., Duke, D., Eustace, J., Galway, A. and Patterson, E. (2008). The humanities hypermedia centre @ Acadia University: An invitation to think about higher education. *Digital Humanities Quarterly* 2(1).
- Dacos, M. (2013). *La stratégie du Sauna finlandais*. Available at: <http://bn.hypotheses.org/11138> (Accessed 11 August 2015).
- Davidson, C.N. (2012). *Now you see it: How technology and brain science will transform schools and business for the 21st century*, reprint ed. New York: Penguin Books.
- Davidson, C.N. (2015). Why yack needs hack (and Vice Versa): From digital humanities digital literacy. In P. Svensson and D.T. Goldberg (eds.), *Between humanities and the digital*. Cambridge, MA: MIT Press, pp. 131–144.
- Davidson, C.N. and Surman, M. (2012). *Why web literacy should be part of every education*. Available at: www.fastcompany.com/1680264/why-web-literacy-should-be-part-of-every-education (Accessed 13 April 2017).
- Drucker, J. (2011). Humanities approaches to interface theory. *Culture Machine* 12. Available at: www.culturemachine.net/index.php/cm/article/view/434 (Accessed 30 April 2016).
- Frank, M., Walker, J., Attard, J. and Tygel, A. (2016). Data literacy – what is it and how can we make it happen? *The Journal of Community Informatics* 12(3). Available at: www.ci-journal.net/index.php/ciej/article/view/1347 (Accessed 15 May 2017).
- Gee, P. (2004). *Situated language and learning: A critique of traditional schooling*. New York: Routledge.
- Gillen, J. (2014). *Digital literacies*, 1 ed. London and New York: Routledge.
- Gilster, P. (1998). *Digital literacy*, new ed. New York: John Wiley & Sons.
- Hafner, C.A., Chik, A. and Jones, R.H. (2015). Guest editor commentary. *Language Learning & Technology* 19(3): 1–7.
- Haigh, T. (2014). We have never been digital. *Communications of the ACM* 57(9): 24–28.
- Hewson, H.C., Vogel, C. and Laurent, D. (2016). *Internet research methods*. 2nd revised ed. Thousand Oaks, CA: Sage Publications Ltd.
- Hughes, L., Constantopoulos, P. and Dallas, C. (2015). Digital methods in the humanities: Understanding and describing their use across the disciplines. In S. Schreibman et al. (eds.), *A new companion to digital humanities*, 2nd revised ed. Chichester, West Sussex: Wiley-Blackwell, pp. 150–170.
- Jenkins, H., Clinton, K., Purushotma, R., Robison, A.J. and Weigel, M. (2009). *Confronting the challenges of participatory culture: Media education for the 21st century*. Available at: www.macfound.org/media/article_pdfs/JENKINS_WHITE_PAPER.PDF (Accessed 15 May 2017).
- Jessop, M. (2005). Teaching, learning and research in final year humanities computing student projects. *Literary and Linguistic Computing* 20(3): 295–311.

- Jones, R.H. and Hafner, C.A. (2012). *Understanding digital literacies: A practical introduction*. Milton Park, Abingdon, Oxon and New York: Routledge.
- Jones-Kavalier, B.R. and Flannigan, S.L. (2006). *Connecting the digital dots: Literacy of the 21st century*. Available at: <http://er.educause.edu/articles/2006/1/connecting-the-digital-dots-literacy-of-the-21st-century> (Accessed 15 May 2017).
- Kirschenbaum, M. (2009). *Hello worlds*. Available at: www.chronicle.com/article/Hello-Worlds/5476/ (Accessed 22 October 2016).
- Kirschenbaum, M.G. (2010). What is digital humanities and what's it doing in English departments? *ADE Bulletin*: 55–61.
- Lankshear, C. and Knobel, M. (2003). *New literacies: Changing knowledge and classroom learning*. Buckingham: Open University Press.
- Lave, J. and Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*, 1st ed. Cambridge and New York: Cambridge University Press.
- Liu, A. (2011). *Where is cultural criticism in the digital humanities?* Blog post. Available at: <http://liu.english.ucsb.edu/where-is-cultural-criticism-in-the-digital-humanities/> (Accessed 16 March 2015).
- Marsh, J. (2007). New literacies and old pedagogies: Recontextualizing rules and practices. *International Journal of Inclusive Education* 11(3): 267–281.
- Mauro, A. (2016). Digital liberal arts and project-based pedagogies. In Constance Crompton et al. (eds.), *Doing digital humanities: Practice, training, research*, reprint ed. New York: Routledge, pp. 373–383.
- Montfort, N. (2016). Exploratory programming. In S. Schreibman et al. (eds.), *A new companion to digital humanities*, 2nd revised ed. Chichester, West Sussex: Wiley-Blackwell, pp. 98–109.
- Mosco, V. (2005). *The digital sublime: Myth, power and cyberspace*, new ed. Cambridge, MA: MIT Press.
- Palmer, C.L., Tefreau, L.C., Pirmann, C.M. and Oclc Research. (2010). *Scholarly information practices in the online environment: Themes from the literature and implications for library service development* [Archived Web Site]. Library of Congress. Available at: www.loc.gov/item/lcwa00095544/.
- Pandey, A.V., Manivannan, A., Nov, O., Satterthwaite, M. and Bertini, E. (2014). The persuasive power of data visualization. *Transactions on Visualization and Computer Graphics, IEEE* 20(12): 2211–2220.
- Pannapacker, W. (2009). The MLA and the digital humanities. *The Chronicle of Higher Education*. Available at: <http://chronicle.com/blogPost/The-MLA-the-Digital/19468/> (Accessed 10 March 2015).
- Prensky, M. (2001). Digital natives, digital immigrants part 1. *On the Horizon* 9(5): 1–6.
- Ramsay, S. (2011a). *On building*. Available at: <http://stephenramsay.us/text/2011/01/11/on-building.html> (Accessed 29 January 2013).
- Ramsay, S. (2011b). *Reading machines toward an algorithmic criticism*. Urbana: University of Illinois Press.
- Ramsay, S. (2011c). *Who's in and who's out*. Available at: <http://stephenramsay.us/text/2011/01/08/whos-in-and-whos-out/> (Accessed 15 May 2017).
- Rockwell, G. and Sinclair, S. (2016). *Hermeneutica: Computer-assisted interpretation in the humanities*. Cambridge, MA: MIT Press.
- Rogers, R. (2013). *Digital methods*. Cambridge, MA: MIT Press.
- Rogers, R. (ed.). (2000). *Preferred placement: Knowledge politics on the web*, 1st ed. Maastricht, The Netherlands: Jan Van Eyck Editions.
- Rushkoff, D. (2011). *Program or be programmed: Ten commands for a digital age*, 1st ed. Berkeley, CA: Soft Skull Press.
- Sanz, A. and Goicoechea, M. (2012). Literary reading rituals and practices on new interfaces. *Literary and Linguistic Computing* 27(3): 331–346.
- Selfe, C.L. (1988). Computers in English departments: The rhetoric of technopower. *ADE Bulletin* 90: 63–67.
- Siemens, R. and Sayers, J. (2015). Toward problem-based modeling in the digital humanities. In P. Svensson and D.T. Goldberg (eds.), *Between humanities and the digital*. Cambridge, MA: MIT Press, pp. 145–162.

- Stone, G. (2011). *The digital literacy classroom*. Leicester: UKLA Publications. Available at: https://ukla.org/shop/details/the_digital_literacy_classroom1 (Accessed 15 May 2017).
- Suárez, J.L. (2013). El Humanista Digital. *Revista de Occidente*: 3805–3822.
- Summey, D.C. (2013). *Developing digital literacies: A framework for professional learning*, 1 ed. Thousand Oaks, CA: Corwin.
- Svensson, P. (2012). The digital humanities as a humanities project. *Arts and Humanities in Higher Education* 11(1–2): 42–60.
- Svensson, P. and Goldberg, D.T. (eds.). (2015). *Between humanities and the digital*. Cambridge, MA: MIT Press.
- Thomas, M., Gee, J. and Palfrey, J. (2014). Foreword. In *Digital skills: Unlocking the information society*, 2014 ed. New York: AIAA.
- Thompson, K.M., Jaeger, P.T., Taylor, N.G., Subramaniam, M. and Bertot, J.C. (2014). *Digital literacy and digital inclusion: Information policy and the public library*. Lanham: Rowman & Littlefield Publishers.
- UNESCO. (2006). *Education for all: A global monitoring report*. Available at: www.unesco.org/education/GMR2006/full/chapt6_eng.pdf (Accessed 24 April 2017).
- Unsworth, J. (2000). *Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this?* Available at: <http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html> (Accessed 3 August 2015).
- van Deursen, A. and van Dijk, J.A.G.M. (2014). *Digital skills: Unlocking the information society*, 2014 ed. New York, NY: AIAA.
- van Zundert, J. and Haentjens Dekker, R. (2017). Code, scholarship, and criticism: When is code scholarship and when is it not? *Digital Scholarship in the Humanities* 32(suppl_1): i121–i133. Available at: <https://doi.org/10.1093/llc/fqx006>.
- Vee, A. (2013). Understanding computer programming as a literacy. *Literacy in Composition Studies* 1(2): 42–64.
- Warwick, C., Terras, M., Huntington, P. and Pappa, N. (2008). If You build it will they come? The LAIRAH study: Quantifying the use of online resources in the arts and humanities through statistical analysis of user log data. *Literary and Linguistic Computing* 23(1): 85–102.
- Weller, M. (2011). *The digital scholar: How technology is transforming scholarly practice*. London: Bloomsbury.
- White, D.S. and Cornu, A.L. (2011). Visitors and residents: A new typology for online engagement. *First Monday* 16(9). Available at: <http://firstmonday.org/ojs/index.php/fm/article/view/3171> (Accessed 10 March 2015).

English language and English literature

New ways of understanding literary language using psycholinguistics

Kathy Conklin and Josephine Guy

Introduction

This chapter assesses the utility of digital humanities approaches to understanding literary language in the context of some methodologies employed in psycholinguistics. The attraction of these technologies for digital humanities researchers into the nature of literariness is in their promise to impart greater rigour to descriptions of literary language. In this way digital humanities methods have been claimed to address some of the widely documented limitations to traditional literary stylistics, which have enabled researchers into literariness from other fields – such as literary critics and literary historians – to be dismissive of linguistic approaches.

For example, in a recent overview of stylistics research Hakemulder and van Peer (2016) voice scepticism about the discipline's attempt to 'free itself from the typical subjectivity that characterizes literary criticism', noting further that 'to say that stylistics has now overcome traditional literary criticism is a form of megalomania'. They question the point of descriptive linguistic analysis, especially of literary texts, in the absence of any evidence that the linguistic features stylisticians typically attend to, such as those associated with foregrounding or defamiliarisation, yield the effects predicated of them on readers. As they bluntly observe: 'most of the conclusions drawn from text analyses, which project them on what could be the reaction of a reader, are merely what they are: projections' (Hakemulder and van Peer 2016: 190–192). Hakemulder and van Peer see the key challenge for stylistics – if it is to have 'a future' – as that of embracing a 'more scientific approach' (what they refer to as 'empirical stylistics'), and suggest that the best way of achieving this goal lies in collaboration between stylisticians and psychologists (Hakemulder and van Peer 2016: 202).¹ Their mention of psychology is significant, in that it associates the need for explanatory rigour with a better understanding of how readers respond to texts, rather than with more detailed analyses of the particular linguistic features or patterns readers are presumed to be responding to when engaged in literary interpretation. In this view, it is only through empirical investigation of actual readers' responses (as opposed to the hypothesised

or idealised reader of traditional reader-response theory) that we gain purchase on the task of identifying which linguistic features are implicated in the generation of literary meaning, and therefore also in determining the extent to which literariness can be assumed to be a property of language-use or, as it has been termed, to be ‘text-driven’ (Stockwell 2016: 151). However, in recent years some of the main advances claimed for this last task have come not from psychology, but from the application of technologies associated with digital humanities research, notably the use of corpus linguistic tools in the analysis of literary texts. We will therefore begin by considering in more detail why digital humanities research appears not to be fulfilling the demand for objectivity invoked by Hakemulder and van Peer.

The use of corpus tools in literary research is wide-ranging, covering activities such as the construction of digital concordances (which have been especially useful in Shakespeare studies), the compilation of searchable digital databases (which have proved invaluable in identifying allusions and in studies of intertextuality more generally, as well as in making fugitive texts more readily available) and the use in digital editing (which we discuss later) of electronic collation programmes such as JUXTA and CASE.² In relation to the claims of literary stylistics, however, the areas of digital research which are of most relevance are those which use corpus tools to gain a better understanding, as we indicated, of the specific linguistic features involved in literary interpretation. Of these, three prominent examples will serve to illustrate the limitations which can attend this sort of investigation: research into genre recognition, into author attribution and into literary characterisation.

The perceived advantage of corpus linguistic tools when applied to these issues is that they can identify very subtle, as well as multiple combinations of, linguistic patterns in large bodies of textual material. Computational models based on the frequency of particular words and structures have thus been used to investigate genre, and machine-generated recognition in this area has been shown, as McCarthy reports, to be both rapid and accurate (McCarthy 2015: 102–103; see also Louwerse et al. 2008), a finding often interpreted as lending support to stylisticians’ assumptions about the text-driven nature of literariness – what McCarthy terms ‘inherent features’ of a text that ‘activate literary responses’ (McCarthy 2015: 102). Computational models have also been used to help settle questions about literary authorship by identifying linguistic patterns, which might involve lexical choices, syntactic constructions, the use of function words and so forth, with authorial signatures. Stylometrics, as it has been termed, is, once again, probably most widely known (if contentious) in Shakespeare scholarship (see, for example, Rudman 2016). In cases like these, where digital analytical methods depend upon statistical probabilities and pattern matching, it is generally the case that the greater the number of variable textual features included in a computational model, the greater the likelihood of accuracy. Yet it is by no means clear that real readers match patterns in the same way as machines: that they respond to, or consciously recognise, either the same kind or number of textual features as those that have been required for effective computer-generated recognition, nor that the cues that readers require for such recognition are necessarily only linguistic. That textual attribution is held to be so important in the first instance, while partly a matter of economics (certainly in relation to the Shakespeare industry), also speaks to the complexity of ‘the author’ as a construct in part brought to, rather than simply derived from, a given text (see, for example, Kirwan 2015). While there are clearly many contexts of literary study in which quantitative evidence for stylistic similarities may be useful and relevant, this kind of machine-generated data tells us very little *on its own* about what constitutes literariness, or which linguistic features, in which kinds of reading situations, are implicated in judgements about literary value.³ (There are additional problems with computer-generated attribution, relating to the algorithms used, the choice

of textual markers, the range of texts sampled and their accuracy; for a useful overview see Koppel et al. 2009.) A more complex and ambitious use of computational methods to investigate literary language can be seen in the CLiC Dickens project (clic.bham.ac.uk; see also Mahlberg 2013) and the use of corpus generated searches to ‘identify textual patterns that function as cues to triggering character information’ (Stockwell and Mahlberg 2015: 129). In this instance, the frequency of certain patterns – word clusters associated with the body – in literary and non-literary corpora are compared to determine those stylistically relevant to literary interpretation, and concordance searches are then conducted to establish how repeated body-word clusters are used to distinguish one character from another. While the findings of this kind of research are undoubtedly interesting, the mere identification of linguistic patterns through digital analysis, even if they *are* more frequent in some sorts of texts as opposed to others (the methodology employed by Stockwell and Mahlberg unfortunately falls short in establishing the higher incidence of body-language clusters in literary as opposed to non-literary texts), still tells us nothing about whether they are salient to the ways in which readers engage with those texts. At best, this use of corpus tools points up data which *may* have a role to play in literary processing, the significance of which will then need to be tested using other methods. Before explaining how psycholinguistics might help us achieve this last aim, there is a further critique of literary stylistics we wish to consider that has arisen from a different area of the discipline of English, and which, as we alluded to earlier, was also an earlier adopter of digital technology: text editing.

At first sight, it might seem that literary stylisticians and text editors should have much in common with each other through their shared attention to the concrete minutiae of language use – or rather, of an identification of literary value *with* that use. The whole rationale of a variorum edition presupposes that the choice of this lexical item rather than that, or the placing of a comma here rather than there, matters to literary judgement. In the editing of modern works, the variorum edition provides a unique window, editors argue, into the workings of a writer’s creative mind.⁴ It is by showing that a writer actually made such choices, through the activity of drafting and redrafting the text of a poem or novel, that variorum editors try to exhibit evidence of literary craftsmanship – a task which has proved important in overthrowing prejudices about a writer’s reputation, especially when (as in the case of figures such as D. H. Lawrence or Oscar Wilde) their works have been understood as ideologically motivated or as having been unduly influenced by commercial pressures. Recently some stylisticians, notably Sotirova (Sotirova 2014), have begun to engage directly with editorial research, developing a descriptive analysis of the textual variants editors document, to shine new light on the nature of authorial creativity. This work notwithstanding, and for reasons we set out later, it remains the case that the views of literariness that have arisen from, or (and perhaps more accurately) have been enabled by, the application of digital technology to traditional editorial practice significantly complicate the research challenge set out by Hakemulder and van Peer and undermine the premises upon which much stylistic research has been based.

Of fundamental importance to modern text editing is the observation that literary works typically survive in multiple textual embodiments and that the linguistic texts of these embodiments may differ from each other in several ways. As Jerome J. McGann (McGann 1991) put it, the defining feature of the process of textual transmission, of what he terms the ‘textual condition’, is ‘change’. In different versions of nineteenth- and early twentieth-century novels, these textual differences may run into the tens or even hundreds of thousands. Modern editors, influenced by the work of Gérard Genette (Genette 1997), have also drawn attention to profound differences in the material and paratextual characteristics of

such embodiments – to features such as epigraphs, prefaces and author biographies, as well as to cover designs, bindings, ink, typefaces, layout, paper quality and so forth. As a result, the modern editor's notion of what constitutes 'the text' of a literary work – and therefore of the kind of data with which she should be concerned when thinking about literariness – has been greatly expanded. It can include many kinds of documents (from pre-publication 'avant-texts', such as notebook jottings, working drafts, fair-copy manuscripts and type-scripts) as well as a whole panoply of non-linguistic features (Guy and Small 2012). This range of evidence poses an intractable problem with regard to pinning down what constitutes 'the (literary) work' – or the object of study – that is to be subjected to linguistic analysis in the attempt to understand what literariness 'is'. Is 'the work' the totality of all surviving embodiments, or is it to be identified with just one of them; if one – then which one, and for whom? There is also the question of which features of that single text – even if it can be agreed upon – 'matter' in attributions of literary value. Where do we derive criteria of relevance for understanding the specific linguistic or material characteristics of a text readers are responding to when making judgements about its literariness? What do we mean when we talk about the literary merit of works like *The Rubáiyát of Omar Khayyám*, for which multiple authorised linguistic and material texts have co-existed for over a century (Guy and Small 2012: 73–85)? The stylistic analysis of elements of any one of these *Rubáiyáts* would, arguably, tell us very little about why it, rather than many other late nineteenth-century literary works with oriental themes, has retained its cultural centrality.

Faced with these dilemmas, and exploiting the potential of digital technology to overcome the constraints of space and display inherent in codex editions, some modern editors have eschewed the traditional task of establishing (or identifying with a work) a single, fixed, authoritative text, in favour of displaying for the reader, in as neutral a manner as possible, the entire body of textual evidence, with a view to allowing readers to become their own editors. Rather than textual variants being arranged in a hierarchy of development and listed separately from the base or copy text in a series of textual notes, digitisation enables the presentation of multiple texts and for textual variants to appear side by side, or as simultaneously possibilities. Realising textual evidence in this manner has led some editors to suggest that literary creativity exists in the principle of variability associated with the textual condition, in what has been termed the 'open' or 'plural' text.⁵ It is not at all clear how the 'scriptable' rather than 'lisible' text described by D. C. Greetham, and operating under the assumptions of poststructuralist 'différance' (Greetham 1993: 17), offers itself for linguistic analysis. The limitation of most literary stylistic analysis in respect of understanding what constitutes literariness, from the point of view of text editors, is a failure to appreciate that 'the text' is as problematic and neglected an object of study, as the readers who respond to it. This in turn is why editors are often sceptical of corpus stylistic analysis, given that it typically works with what they consider inaccurate or non-authorised digitised texts – such as those produced by Project Gutenberg (www.gutenberg.org).⁶ In short, the main message of text editing over the last three decades, underlined by the use of digital technology, is that there is no single, stable object of study – 'the literary text' – to be subjected to linguistic analysis. And this is why, for many text editors, literary stylistics (whether in traditional or digital manifestations) is too often seen as dodging the questions about literariness which editorial research has made central, namely: the extent to which literariness can indeed be assumed to be an inherent feature or property of the linguistic text, as opposed to deriving (at least in part) from the specific material form(s) by which we encounter it; and, given that literary works typically exist in plural embodiments, which may differ linguistically in many ways, how do we establish which stylistic features are key in literary evaluation?

It is exactly these sorts of issues which, as we will now go on to explain, psycholinguistic methodologies can help us address.

Current contributions and research

Psycholinguistics offers a reader-centred investigation of language use which seeks to draw generalisable conclusions from data taken from ‘real’ readers and their reading ‘experiences’. Unlike most current literary linguistic approaches, it arrives at those conclusions through controlled experiments: the focus is on how various linguistic phenomena influence *actual* reading behaviour – not hypothesised reading behaviour. A key concern for literary stylisticians of all types is the analysis of linguistic features at various levels (phonological, morphological, lexical, syntactic and pragmatic) to consider how the rules of language are either observed or violated to create specific effects within a text (Simpson 2004), which in turn may have broader ideological implications (Stockwell 2005). Some literary linguists explicitly claim that a close linguistic analysis can tell us about ‘reader experience’ in the sense that it relates to specific effects in the minds of readers (Gavins and Steen 2003). Crucially, the ‘reader’, in this kind of argument, is typically treated as an entirely abstract entity; as we noted, arguments are made about the presumed effect that a linguistic feature or set of features has without actually investigating it with regard to real readers and authentic texts. When literary linguists do collect data from readers, participants are generally asked to reflect on their thought processes while reading a text, which interrupts the largely automatic ‘online’ behaviour of reading. Responses may also be elicited shortly after the reading has been completed through ‘think-aloud’ protocols, questionnaires, surveys and/or various kinds of rating scales – all of which are ‘off-line’ activities that involve a conscious, *post hoc* reflection on what was experienced *at the time* as part of a largely unconscious set of behaviours.

This kind of empirical research, which relies on readers’ verbal reports, is widely used in other areas of applied linguistics (such as second language learning) and can provide useful data which can be analysed to identify common attitudes or patterns of behaviour among selected groups of readers. However, its value in understanding cognition is problematic, as there is no guarantee that what readers say they were doing or experiencing while reading is what they were actually doing or experiencing. Corbett, for example, cautions against the ‘easy equation between what readers say they are doing and the actual cognitive processes which are occurring as they read’ (Corbett 1997: 199), although he goes on to suggest that this approach can still be beneficial as a way of connecting the text with the people experiencing it. Others have expressed concerns about the reliability of readers’ memory and about the possibility that the ways in which readers are asked to articulate, categorise or rate their experience may shape what they report, whether by ‘contaminating’ it or, more positively perhaps, by modifying it in a ‘category-changing’ manner. In this respect, the significance of psycholinguistic methods is that they focus on readers’ *unconscious* processes, through measurements of eye movements or forms of neural activity, delivering what is commonly referred to as ‘behavioural’ rather than ‘experiential’ data.⁷ The relationship between behavioural and experiential data in empirical investigations of literariness is currently a matter of controversy.

Some commentators have objected to what they see as the privileged status afforded to behavioural data on the grounds that it is assumed to be more objective or scientific than information gleaned from readers’ subjective reports. Thus, Kuiken has pointed to the difficulties of designing experimental studies which avoid the problem of ‘reverse inference’

(Kuiken 2016: 172),⁸ questioning the role of behavioural data per se in investigating a phenomenon – literary experience – which is made up of so many variables. These may include factors such as education and prior exposure to similar kinds of texts (implicit in understanding genre), as well as mood and general ‘life’ experiences (a reader who has been the victim of rape may view the famous bowdlerised passages from *Tess of the D’Urbervilles* differently from one who has not). This complex, multifaceted process is not the same as, and cannot easily be reduced to, aspects of a local ‘response’ measured, in whatever ways, in a laboratory. Other behavioural sceptics have gone so far as to argue that the highly unnatural conditions of the lab also need to be viewed as a potential source of contamination, which may be no less distorting than reader subjectivity, insofar as these conditions may induce mental processes that would not have occurred in the absence of the activity of measuring (Dixon and Bortolussi 2016). A further unnatural feature of the experimental setting is that, and with the exception of fairly short poems, it is difficult to measure a reader’s experience of an entire text. Reading a nineteenth-century realist novel will typically take place in a series of episodes over days or even weeks; even reading relatively short pieces of poetry or prose may involve a process of reflection that lasts, as Dixon and Bortolussi explain, ‘long after one finishes moving one’s eyes over the words of the text’ (Dixon and Bortolussi 2016: 178). It is these kinds of reservations that have led some critics to conclude that a behavioural understanding of what happens when readers experience literary language, or the text of a literary work, is ‘technically impossible’ (Willems and Jacobs 2016: 243). Recent research, however, is challenging such suppositions, as researchers have become more adept in mimicking natural reading contexts in a laboratory setting. This is exemplified by a series of publications that has resulted from work by Cop and her colleagues in which participants were asked to read an entire Agatha Christie novel, *The Mysterious Affair at Styles* (Dutch Title: *De zaak Styles*) while their eye movements were monitored (Cop et al. 2015, 2016; Cop, Dirix et al. 2016; Cop, Keuleers et al. 2015). Participants, monolingual English speakers and native Dutch speakers, read the novel in four sessions of an hour and a half each: in the first session participants read chapters 1 to 4, in the second 5 to 7, in the third 8 to 10 and in the fourth 11 to 13. The monolinguals read the book entirely in English, while the non-native participants read chapters 1 to 7 in one language and 8 to 13 in the other. Crucially, this type of research gives a rich data source that can be explored in numerous ways, while at the same time challenging the view that it is impossible to measure a reader’s experience of an entire text over a series of episodes.

Unfortunately, it is beyond the scope of the present chapter to provide a comprehensive analysis of the relative validity of experiential versus behavioural data; suffice to say, there are merits and limitations to *both* kinds of approaches, but for the study of literature to be considered ‘scientific’, it *must*, as Jacobs observes, involve investigating both ‘the conscious aspects of, and the unconscious processes underlying, literary experience’ (Jacobs 2016: 143). Our contention is that in distinction to much of literary stylistics, including that using corpus linguistic tools, psycholinguistic methods are currently the only approach which can help meet this demand. And that while it remains very challenging, advances in technology and methodology are making such a scientific understanding increasingly less elusive. So long as caution is exercised in the use of behavioural data, there is, as even Kuiken concedes, a ‘rich range of opportunities for empirical studies of literary reading’ (Kuiken 2016: 171). For our purposes, these opportunities lie not in claiming to be able to explain literary experience by reference to something called ‘literary processing’, but in determining which aspects of the unconscious mental processing that underlies the experience of reading a literary text are measurable via psycholinguistic methods, as well as, importantly,

the disciplinary debates about literariness in which these measurements, or data, may have relevance. We therefore turn next to describing these methods in more detail before proceeding to discuss how they can be applied to investigating some literary problems.

Of central importance to psychology and psycholinguistics is the concept of mental processing, understood by reference to various sub-conscious physiological reactions, such as eye movements, skin conductance, muscle movement and activity in the brain – all of which can be measured under experimental conditions via eye tracking, skin sensors, electroencephalography (EEG) and functional magnetic resonance imaging (fMRI). Psycholinguists are interested in the mental processes that underpin language comprehension, and which they aim to understand by correlating these behavioural or neural measurements with controlled uses of, or responses to, language (whether spoken or read) in a laboratory setting. A primary concern of designing any psychoanalytic study of language is *ecological validity* – or to what extent the processes under investigation also occur under ‘normal’ conditions (Zwaan 1993). Psycholinguists generally prepare carefully controlled and manipulated texts; and it is these texts which form their object of study. Thus, a participant in a psycholinguistic study may encounter 100 sentences that are all the same length; have a similar syntactic complexity; and which contain words of a particular frequency, imaginability, concreteness, etc. This level of control is necessary because it allows researchers to attribute the locus of an effect. Because every element of the experiment is the same, other than the variable that is being studied and manipulated, any difference in participant behaviour can be attributed to the variable of interest. While this type of experiment might tell us about a very particular element of language processing, most would not consider it ecologically valid, as it differs significantly from a natural reading experience. Ecological validity is exceptionally challenging when investigating the processing of literary works, for reasons discussed earlier. However, studies can be devised which have more (or less) ecological validity.

For example, in the research described in an earlier section, the texts that readers encounter are all extracts from actual versions of literary works – they are not texts that have been created or modified simply for experimental purposes, a situation which is at odds with general practice in psycholinguistics.⁹ The tension between the need to present carefully controlled materials and to ensure a reader’s experience is as authentic as possible can also be addressed by the use of analysis techniques like mixed-effects modelling that allow researchers to consider variables, such as word frequency, imaginability, etc., that are not controlled experimentally. Another methodological concern when investigating the processing of authentic works relates to the number of experimental items (Zwaan 1993). When researchers develop their own stimuli, they can create sufficient numbers of them to ensure that the question under investigation can be assessed statistically. Researchers using authentic literary texts are reliant on the number of occurrences of a phenomenon in the text itself. They need to evaluate whether there are enough instances of the phenomenon to make it testable. If the answer is ‘no’, they need to decide whether artificially creating additional examples is possible and if this is worth a sacrifice of ecological validity.

Possibly the main challenge to ecological validity when investigating the processing of literary works is, as noted earlier, the laboratory environment itself. Participants know that they are in a lab and are aware that at least some element of their behaviour is under investigation. Different experimental techniques and paradigms may make this easier or harder for participants to ignore. Common psycholinguistic technologies involve applying electrodes to the skin and/or scalp: electrodermal activity (EDA), which measures increases in skin conductance that are caused by sweating due to emotionally arousing stimuli (Dawson et al. 2007); facial electromyography (EMG), which detects the facial muscle movement that

indicate the presence and degree of certain emotional responses like frowning or smiling (van Boxtel 2010); and EEGs and event-related potentials (ERP) that measure the electrical activity of the brain, which in ERP is related to a particular linguistic event, such as the appearance of a particular word (Harley 2014). While EDA, EMG and EEG/ERP are all non-invasive technologies, they can have long and cumbersome preparation procedures that can cause some discomfort to participants because of the wet, gel-covered electrodes (Grübler et al. 2014). Another non-invasive technique is fMRI, which in recent years has become more accessible to researchers and is quickly turning into ‘one of the most important sources for data in psychology’ (Harley 2014: 20). Nonetheless it can also cause discomfort and distress due to lying prone in the close confines of a scanner, as well as because of acoustic noise (Kulynych 2002). Further, for many of these methods, in order to localise the precise part of a text (a word or phrase) that elicits an effect, presentation of the text itself is often quite unnatural. In the case of ERP, the words of a text are generally presented one at a time on the middle of a computer screen for 200 to 300 ms (ms is a millisecond, or a thousandth of a second) followed by a blank screen for 300 ms. This helps to ensure that the electrophysiological signal generated by a word is not contaminated by artefacts from eye movements, and that the signal can be attributed to a specific word. While EDA, EMG, EEG, ERP and fMRI all allow for sections of text larger than an individual word to be presented, doing so makes it difficult to pinpoint the exact origin of any effect.

In contrast to the technologies just described, eye tracking provides a context that is most akin to natural reading, as it involves reading text from a computer screen without the need for any behavioural response (e.g. button press). By providing a rich moment-to-moment record of eye movements, eye tracking affords an understanding of how features such as word choice, syntactic structures and other identifiable patterns affect the online (unconscious) processing of texts in actual readers. This technology assumes an ‘eye-mind equivalence’, whereby what is being looked at is thought to be what is being processed at any given time (Pickering et al. 2004). This means that the eyes indicate what is being paid attention to and how much cognitive effort is being expended to process the input. Specifically, the technology tells us where people’s eyes land, how many times they land in that position or region (*fixation count*) and how long each fixation lasts (*fixation duration*), as well as measuring the movement of the eyes (*saccades*) in terms of duration and length. Thus, fixations provide a measure of processing effort; longer and/or greater fixations provide an indication of greater processing effort or attention, which can be attributed to (sub-conscious) noticing. It is worth noting that ‘effort’, in this context, does not distinguish between the kind of difficulty associated with unfamiliar or novel linguistic constructions (such as might be associated with defamiliarisation) and a lingering which may be due to the pleasure of experiencing a given linguistic feature. However, these distinctions may be picked up through combining eye tracking with readers’ conscious reports of their experience; while eye tracking itself, by furnishing a measure of pupil dilation, can give an indication of other kinds of emotional response when reading (Laeng et al. 2012): fear, arousal, anticipation, risk, novelty, surprise and conflict all cause pupil dilation to increase.

In theory, it is possible for eye tracking to be used to investigate large portions of text – such as a whole poem – as well as aspects of the visual field implicated in McGann’s ‘material text’, such as the significance of illustrations, different fonts and ink colour (McGann 1991).¹⁰ The technology has been employed, for example, better to understand how readers encounter concrete poetry (Shingler 2011). Using eye tracking to understand fine-grained elements of literary interpretation, such as discerning the *relative* semantic significance of various linguistic or visual elements of a text, is still constrained by the limitations of the

technology. Eye trackers are less accurate at measuring vertical than horizontal movement. Even when concentrating only on the linguistic text, double-line spacing, or even triple-line spacing, is recommended to ensure that fixations can be reliably attributed to one line or another (Conklin et al. 2018). By the same token, reliable interpretations of non-linear eye-movements – such as, for example, the movement between text and image in an illustrated poem by William Blake – remain problematic, unless these are relatively coarse.

An important methodological concern in scientific research is that a study and its findings should be replicable. To achieve this, other researchers need to be able to access the materials used in a study. When researchers create their own stimuli, this is straightforward; they can be made available in an appendix or upon request. When using published materials from a literary text, there may be copyright issues with providing them freely. Crucially, so that others can replicate and/or extend the research it is imperative that the source of the materials is clear. As was highlighted in the previous section, versions of a single literary work may contain numerous textual differences: in the examples that follow, these may be in lexis (1) and/or in punctuation (2).

(1) Henry James, *Confidence*

MS (manuscript)	“and I find your apartment so charming.”
1879 (English book)	“and I find your apartment enchanting.”
1880 (American book)	“and I find your apartment magnificent.”
SM (American serial)	“and I find your apartment charming.”

(2) Henry James’s *Confidence*

MS (manuscript)	than for the button on her glove: by which I mean
1879 (English book)	than for the button on her glove; by which I mean
1880 (American book)	than for the button on her glove – by which I mean
SM (American serial)	than for the button on her glove, by which I mean

As we will see in the following section, such differences may be noticed by readers and influence their processing. Thus, the choice of one version of a literary work over another is not a trivial question. When using psycholinguistic methods to investigate the processing of literary texts, researchers must make a number of decisions about the most appropriate methods to answer their research question. This will involve deciding on the most suitable materials, participants, procedure, experimental technique, analyses, etc. When making these decisions, it is important to keep in mind how each choice impacts both the ecological validity and replicability of a study.

Another important aspect of applying psycholinguistic methods to the study of literary works involves getting into a ‘scientific’ mind-set. In his paper in *Physics and Reality*, Albert Einstein tells us that ‘the whole of science is nothing more than a refinement of everyday thinking’ (Einstein 1936: 349). This refinement can come about through the application of the scientific method, which involves observation, hypothesis generation, testing and finally hypothesis rejection/revision, or alternatively development of a theory. In the examples in (1) and (2) this method would start with the observation that the manuscript and three published editions of Henry James’s *Confidence* differ in terms of lexical choices and punctuation. A reasonable hypothesis might be that if fine details of language are an important element of literariness, as is widely assumed, then readers should notice these

differences. The part of this hypothesis that can be tested in an experiment is the ‘noticing’: if it is demonstrated that readers do not register these differences, then it can be reasonably concluded that they likely have little relevance to literary interpretation (at least under the conditions of reading in the experiment). On the other hand, showing that readers *do* pay attention to them only indicates their possible (but not necessary) implication in literary appreciation.¹¹ The value of such an experiment for the literary critic thus lies not in a claim to have explained what literariness ‘is’ by correlating linguistic features with alleged mental phenomena; but in having helped to narrow down the range of textual features to which critics need to attend. As we will explain later, in certain contexts, this apparently modest task can have far-reaching implications.

Sample analysis

When attempting to rate sensitivity to, or examining the processing of, literary works, most empirical studies to date have focused on examining a set of textual features that contribute to what literary linguists refer to as foregrounding. Typically, these have involved an examination of phonetic features (but also sometimes syntactic or even graphic features), such as rhyme schemes and enjambment, and have taken their examples predominantly from poetry. In contrast to this more established research, we have been exploring whether readers are sensitive to quite small changes in textual features in prose fiction, and which have not generally been implicated in foregrounding (linguists have, however, looked at punctuation effects in poetry; see Gómez-Jiménez 2017). At the same time, these small changes are of considerable interest to literary critics and text editors. They are typically the most numerous kinds of changes involved in the processes of textual transmission described by McGann (McGann 1991) and have been afforded significant semantic weight in accounts of the textual condition of certain literary works (see, for example, McKenzie 1986; Frankel 2011). Assumptions about their interpretative value have also underwritten the commissioning of several large-scale literary-editing projects, although there are deep disagreements among individual editors as to the relative importance of small-scale changes or variants, and therefore which of them are worth the editor recording (Guy et al. 2016). Our experiments are designed to help literary critics and text editors address this last issue, by providing them with a better understanding of the contexts, and of the kinds of readers, for which these changes might ‘matter’.

To establish whether readers are sensitive to small-scale changes to punctuation, we presented participants with sentences containing textual variants, like those exemplified in (1) and (2). This was done across three separate studies; the results of Study 1 are described in detail in Carrol et al. (2016) and Studies 2 and 3 in Parente, Conklin, Guy, Carrol and Scott (2020). In all three studies participants were shown extracts from the 1846 and 1867 editions of Charles Dickens’s *Oliver Twist* and from the 1881 and 1908 editions of Henry James’s *The Portrait of a Lady*, with all texts being presented as authoritative.¹² Each experimental item consisted of a pair of sentences that differed in some way between the two editions in terms of lexis (word choice or word order) and/or punctuation (choice of punctuation mark or inclusion/deletion). Filler items from the texts were included that had no change. Participants saw both versions of each item, counterbalanced across two presentation lists in terms of which version was presented first. For the critical items, specific regions of interest (ROIs) were identified which represented the parts of the sentence that differed from the first presentation to the second (e.g. ROI: *evenly distributed features* compared to ROI: *features evenly distributed*; ROI: *hopes: cherished* compared to ROI: *hopes, cherished*). In

Studies 1 and 2, the paired sentences were presented one after the other; but with two main differences. First, Study 1 included sentences that had more than one difference across the two versions, while Study 2 only presented sentences that had a single difference in the pairs. Second, all the participants in Study 1 were native speakers of English studying for undergraduate or postgraduate degrees, although in a variety of subjects. In Study 2, three separate groups of native English speakers were tested: undergraduate students studying English; postgraduate students studying English; and undergraduate students in science, technology, engineering and mathematics (STEM subjects). Our aim in Study 2 was to see whether readers' expertise (a topic widely investigated in empirical studies of literature) had a role to play in the noticing of small-scale variants, as might be expected, given findings of earlier research on the importance of expertise in the generation of literary meaning (see for example, Vipond and Hunt 1984; Graves and Frederiksen 1991; Zeitz 1994; Bortolussi and Dixon 1996; Peskin 2007; Levine and Horton 2013). Study 3 used the same sentence materials as Study 2 (only one difference per sentence pair) and tested English undergraduate students studying English. The main manipulation in this last study was that sentence variants were not presented one after another, but instead in two halves of the experiment; in the first half, one version of all of the sentences were presented one at a time; and after a short break, the second version of all the sentences were presented. As with the first two studies, the sentences in Study 3 were counterbalanced across lists in terms of which version was presented first. The aim of this last study was to see whether readers' noticing had been affected by a 'spot the difference' task-bias in the other two experiments. In all three of the studies, after seeing the second version, readers were asked if they noticed any differences in the two presentations of the sentences, and, if so, whether they could explicitly identify what those differences were.

Across all three of the studies, the results demonstrated that when reading variants of a text with changes to either punctuation or lexis, there was an increase in reading times to the ROI containing the change relative to the rest of the sentence that remained unchanged. There was no evidence in the reading record to suggest that changes to punctuation were less noticeable than changes to lexical items or word order; in fact, Studies 2 and 3 indicated that ROIs with punctuation changes elicited greater processing effort. Importantly, the findings from across the three studies show that all readers, even non-expert ones, pay a certain amount of attention to minor textual features such as the presence or absence of a comma, or the change from a semicolon to a colon. Interestingly, in Study 1, changes in lexis triggered more re-reading of the whole sentence, whereas changes to punctuation did not. This indicates that readers may implicitly ascribe more 'semantic load' to lexical changes, which causes them to reconsider the sentence as well as the change itself. In contrast, noticing changes to punctuation did not seem to cause readers to reconsider the broader sentence, suggesting that such features are deemed minor variations with limited interpretative significance. Studies 2 and 3 showed a slightly different pattern of results. In these two studies, where there was only one change in the sentence (either punctuation or lexical), reading times slowed for the ROI, as in Study 1; however, this had no impact on the reading of the rest of the sentence, meaning that lexical changes did not induce more rereading. Of interest is that in both Studies 2 and 3, there was a greater slowdown in ROIs containing punctuation changes than lexical ones. This indicates that identifying small changes in punctuation is more effortful, and aligns with the participants' ability to consciously identify changes in punctuation, with punctuation changes being identified significantly less accurately. Finally, the primary manipulation in Study 2 demonstrated similar reading patterns for punctuation and lexical changes for the three groups of readers, with the main group difference being,

that with greater experience of reading literary texts in general, and those by Dickens and James in particular, reading speed increases.

How do the findings of these studies help text editors to decide which textual changes are important? Most obviously, editors should be reassured that readers from a wide variety of backgrounds, and not just those well-practised in close reading, seem reasonably adept at noticing quite small textual variants, such as fine-grained details of punctuation. However, less reassuring may be the finding (from Study 1) that noticing a textual feature does not necessarily mean ascribing it interpretative significance. That readers were less likely to re-read a sentence when it contained a punctuation change, and that as all three studies show, they were less accurate in identifying punctuation changes, hints that this sort of textual feature may have less relevance to judgements about literary value than some textual theorists have supposed, even for expert readers supposedly trained to look at such detail. Perhaps the systematic and very time-consuming recording of these features by some text editors is not justified.¹³ However, it still might be wondered why, in Studies 2 and 3, the re-reading effect noticed in Study 1 with respect to lexical items did not occur. This may be because the processing effort of noticing one sort of change at a time, particularly when it is relatively obvious (like a lexical item), is sufficiently easy, so as to enable readers to take account of any interpretative relevance without the necessity of reviewing the sentence. Or it may be that in realising that there was only one change to spot each time, participants in Studies 2 and 3 quickly became focused on the activity of spotting changes, rather than on thinking about their significance. Here it is perhaps relevant that in Study 1, participants were not told in advance how many differences there might be between sentence pairs; thus, it may have been more difficult for them to have judged what was at stake in the test. The re-reading in Study 1 may have been driven by an element of unpredictability in the task, as well as a need (arising from that unpredictability) to make sense of – in effect, give semantic priority to – multiple changes.¹⁴ But that there was more re-reading triggered by the noticing of lexical rather than punctuation items is still highly significant, given that – as we see from Studies 2 and 3 – punctuation items are, in themselves, more effortful to notice. Looking again at the results of Study 1, in light of those of Studies 2 and 3, it is striking that the processing effort expended in noticing punctuation changes, and which all readers exhibit, seems to have such a relatively small semantic pay-off, compared to noticing changes to lexical items.

The methodology and the experiments summarised here clearly do not resolve current disputes in the text-editing literature about the value of recording all types of textual variant; nor do they tell us exactly what role small-scale textual features play in literary appreciation.¹⁵ However, they do demonstrate the potential of empirical data to shift the terms of such debates, by providing editors and critics with some new ways of thinking about the semantic significance of minor textual changes.

Future directions

Real reading is still the noblest of the arts, the medium by which there still come to us the loftiest inspirations, the highest ideals, the purest feelings that have been allowed mankind, – a God-gift indeed, this written word and the power to interpret it. And reading itself, as a psycho-physiological process, is almost as good as a miracle. To the average reader the process by which he gets his pages read is not understood.

(Huey 1908: 5)

Edmund Burke Huey, one of America's first experimental psychologists and the first to apply the scientific method to the study of reading, in his influential book *The Mysteries and Problems of Reading* (1908), rightly points out that the average reader does not understand how she reads. However, advances in technology, as well as in statistical analyses, over the last century have provided us with the means to apply the scientific method to reading so that we can understand elements of this 'miracle'. Technologies like eye tracking, EDA, EMG, EEG, ERP and fMRI have allowed us to begin to unravel the complex set of unconscious physiological responses – such as eye movements, skin conductance, muscle movement and activity in the brain – that occur when reading literary works and which can be measured under experimental conditions. Incorporating these technologies into the arsenal of tools used to study the reading of literature has huge potential, as the evidence they provide is generally not subject to the strategic effects, deliberate decision processes or recognition strategies that are problematic for many types of experiential research data. However, as with any set of new methodologies, a psycholinguistic approach has limitations and constraints. It is not a magic potion that can reveal everything about how the brain works when reading, let alone when reading literary texts. It simply gives us a new set of tools which have certain properties and which can provide specific kinds of information about the mental processes implicated in the reading of authentic texts. The value of this information in informing our understanding of literariness will always be dependent on the good design of psycholinguistic experiments, as well as on researchers' responsiveness to debates about the nature of literariness within the wider discipline, and most particularly and as outlined here, within text editing.

Notes

- 1 This move towards scientism, and most particularly the influence on literary study of linguistics and psychology, is not welcomed everywhere; in a provocative critique of the use of digital technology in the humanities, Sharon O'Dair equates such methodological ambitions with 'neoliberalism' and with the closing down of critical dialogue and the pluralism of critical methods that currently characterise academic literary studies (O'Dair 2014: 127). Her polemic appears in an engaging collection of essays exploring the impact of digital humanities methodologies, with a focus on Shakespeare research, which is admirably even-handed in discussing both the advantages and limitations of such tools (see Carson and Kirwan 2014).
- 2 Digital editing tools have limitations relating to the number of variant texts which can be compared at any one time, the requirement for comparative texts to be digitised and the consequent inability to deal with manuscript materials which often make up a significant proportion of the text editor's data.
- 3 Even very sophisticated models, such as that developed by Arthur Jacobs and his co-collaborators that successfully categorised 154 Shakespeare sonnets into 'young man' and 'dark lady' poems using 96 variables, tell us little about the numbers and kinds of textual features which actual readers require to make such judgements (see Jacobs 2017). As Jacobs et al. admit, using so many variables may involve 'overfitting', meaning that they are 'using too many and thus potentially irrelevant predictors/parameters and thus picking up noise in the data together with the signal' (35). Notably, they demonstrate that their algorithm does well with far fewer variables (between three and ten); however, it is not clear that actual readers would (accurately) track even this small number of variables across 154 sonnets.
- 4 Variorum editing of works belonging to earlier periods tends to be concerned with establishing lines of transmission and determining how certain texts come to be established as authoritative, rather than with the creative agency of any particular individual or individuals.
- 5 This position is most closely associated with genetic editing and genetic criticism, but it also underwrites the case for many electronic editions (see, for example, Shillingsburg 1996).
- 6 Other digital databases such as LION have greater functionality and allow for more systematic checking.

- 7 The most effective stylistic attribution research also focusses on less conscious elements of language use, such as pause patterns, as these are less open to plagiarism or imitation. (We are grateful to Peter Kirwan for sharing with us his knowledge of stylistics.)
- 8 This means being able to identify a specific pattern of unconscious response – such as neural activation – which, as Kuiken puts it, ‘is present in the presence of a cognitive process and also absent in the absence of that cognitive process’ (2016: 172). Kuiken argues that the latter is an especially difficult problem in neuroscience because ‘even quite specific patterns of neural activation regularly mediate several cognitive processes’ (173). He sees the main challenge of behavioural approaches as that of demonstrating convergent *and* discriminant validity.
- 9 A separate issue, which we do not discuss here but which is of interest to psycholinguistics as a discipline, is the extent to which psycholinguistic research into literature has methodological implications for general psycholinguistic inquiry.
- 10 But not, of course, the sensory properties of books as physical objects (i.e. the tactile quality of, say, handmade paper or embossed type). However, other kinds of technologies are being used to study the general effects on reading performance of different media. Research into the visual ergonomics of tablets and e-readers, as opposed to books, suggests engaging with texts via rough or smooth surfaces may affect empathy levels. As Burke et al. (2016) observe, these sorts of results are of ‘pronounced importance’ when reading literature, as books ‘are often seen to have a warm, organic feel to them, whereas electronic devices can feel cold and detached’ (32).
- 11 Establishing that fine details of language are a necessary element of literary appreciation would require a further set of experiments testing whether the attention readers paid to, say, changes in punctuation was the same when experiencing texts not categorised as literary.
- 12 In practice, changes to punctuation may be introduced by the author or by the publisher (with or without the author’s permission), although it is not always possible to be certain which kind of agency is at work. For the purposes of our experiments, and to minimise the number of variables being tested, all textual variants were presented as having the same authority (although we recognise that information about the authority of a given textual change may affect the attention paid to it).
- 13 Here it is worth stressing that the whole rationale of *literary* editing – that is, the editing of a literary versus a non-literary text – is predicated on the assumption that textual authority and textual changes, even fine-grained ones, matter.
- 14 In none of the studies were readers told how many changes there might be in any one sentence pair; nonetheless, it is likely that in Studies 2 and 3, they quickly became aware that there was only one change that needed to be identified.
- 15 To determine that role, further studies would need to be undertaken to ascertain the kinds of conditions under which minor textual features were more or less likely to be noticed by readers. These might include, for example, and as suggested in endnote 12, readers being given information about the authority of a variant; the authors are currently engaged in such studies.

Further reading

- 1 Conklin, K., Pellicer-Sánchez, A. and Carrol, G. (2018). *An introduction to eye-tracking: A guide for applied linguistics research*. Cambridge: Cambridge University Press.
This book introduces eye tracking in an understandable, accessible and achievable manner for language and literature researchers. It explains the technology and what measuring eye movements tells us in traditional experimental contexts, as well as when authentic texts are used.
- 2 Carrol, G., Conklin, K., Guy, J. and Scott, R. (2016). Processing punctuation in different versions of prose fiction. *Scientific Study of Literature* 5(2): 200–228.
This paper presents a study that uses eye tracking to determine how far readers engage with textual variants – small-scale changes to punctuation and lexical changes – and how eye-movement behaviour relates to readers’ conscious ability to report such changes. The authors discuss how eye tracking might be applied to questions of ‘literary’ significance and textual processing.
- 3 Jacobs, A.M. (2016). The scientific study of literary experience. *Scientific Study of Literature* 5(2): 139–170.

- 4 Dixon, P. and Bortolussi, M. (2016). Measuring literary experience: Comment on Jacobs. *Scientific Study of Literature* 5(2): 178–182.
In his paper Jacobs outlines a methodological and theoretical agenda that integrates research and researchers from the humanities who use phenomenological approaches and techniques of text analysis and interpretation with those from cognitive neuroscience whose employ (neuro-) psychological methods and models. The Dixon and Bortolussi paper is a response to this approach.
- 5 Willems, R.M. and Jacobs, A.M. (2016). Caring about Dostoyevsky: The untapped potential of studying literature. *Trends in Cognitive Sciences* 20(4): 243–245.
In this paper, Willems and Jacobs demonstrate that the neurocognitive study of literature is viable from a technical standpoint. They discuss four subdisciplines of cognitive science for which the study of fiction is relevant and that have provided interesting insights.

References

- Bortolussi, M. and Dixon, P. (1996). The effects of formal literary training on literary reception. *Poetics* 23: 471–487.
- Burke, M., Kumičová, A., Mangen, A. and Schilhab, T. (2016). Empathy at the confluence of neuroscience and empirical literary studies. *Scientific Study of Literature* 6(1): 6–41.
- Carrol, G., Conklin, K., Guy, J. and Scott, R. (2016). Processing punctuation in different versions of prose fiction. *Scientific Study of Literature* 5(2): 200–228.
- Carson, C. and Kirwan, P. (eds.). (2014). *Shakespeare and the digital world*. Cambridge: Cambridge University Press.
- CLiC Dickens. Available at: www.birmingham.ac.uk/schools/edacs/departments/englishlanguage/research/projects/clic/index.aspx.
- Conklin, K., Pellicer-Sánchez, A. and Carrol, G. (2018). *An introduction to eye-tracking: A guide for applied linguistics research*. Cambridge: Cambridge University Press.
- Cop, U., Dirix, N., Drieghe, D. and Duyck, W. (2016). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*: 1–14. Available Online First.
- Cop, U., Dirix, N., Van Assche, E., Drieghe, D. and Duyck, W. (2016). Reading a book in one or two languages? An eye movement study of cognate facilitation in L1 and L2 reading. *Bilingualism: Language and Cognition*: 1–23. Available Online First.
- Cop, U., Drieghe, D. and Duyck, W. (2015). Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. *PloS One* 10(8): 1–38.
- Cop, U., Keuleers, E., Drieghe, D. and Duyck, W. (2015). Frequency effects in monolingual and bilingual natural reading. *Psychonomic Bulletin and Review* 22(5): 1216–1234.
- Corbett, J. (1997). *Language and Scottish literature*. Edinburgh: Edinburgh University Press.
- Dawson, M.E., Schell, A.M. and Filion, D.L. (2007). The electrodermal system. In J.T. Cacioppo, L.G. Tassinary and G.G. Berntson (eds.), *Handbook of psychophysiology*, 3rd ed. New York, NY: Cambridge University Press, pp. 159–181.
- Dixon, P. and Bortolussi, M. (2016). Measuring literary experience: Comment on Jacobs. *Scientific Study of Literature* 5(2): 178–182.
- Einstein, A. (1936). Physics and reality. *Journal of the Franklin Institute* 221(3): 349–382.
- Frankel, N. (ed.). (2011). *Oscar Wilde: The picture of Dorian Gray: An annotated, uncensored edition*. Cambridge, MA: Harvard University Press.
- Gavins, J. and Steen, G. (eds.). (2003). *Cognitive poetics in practice*. London: Routledge, pp. 159–181.
- Genette, G. (1997). *Paratexts: Thresholds of Interpretation* (Jane E. Lewin, trans.). Cambridge: Cambridge University Press.
- Gómez-Jiménez, E.M. (2017). Unconventional patterns in the experimental poetry of E. E. Cummings: A stylistic approach to punctuation marks. *Language and Literature* 26(3): 191–212.
- Graves, B. and Frederiksen, C.H. (1991). Literary expertise in the description of fictional narrative. *Poetics* 20: 1–26.

- Greetham, D.C. (1993). Editorial and critical theory: From modernism to postmodernism. In G. Bornstein and R.G. Williams (eds.), *Palimpsest: Editorial theory in the humanities*. Ann Arbor: University of Michigan Press, pp. 9–28.
- Grübler, G., Al-Khodairy, A., Leeb, R., Pisotta, I., Riccio, A., Rohm, M. and Hildt, E. (2014). Psychosocial and ethical aspects in non-invasive EEG-based BCI research – a survey among BCI users and BCI professionals. *Neuroethics* 7(1): 29–41.
- Guy, J., Scott, R., Conklin, K. and Carrol, G. (2016). Challenges in editing late nineteenth- and early twentieth-century prose fiction: What is “editorial completeness”? *English Literature in Transition* 59(4): 435–455.
- Guy, J. and Small, I. (2012). *The textual condition of nineteenth-century literature*. London and New York: Routledge.
- Hakemulder, F. and van Peer, W. (2016). Empirical stylistics. In *The Bloomsbury companion to stylistics*. London: Bloomsbury, pp. 189–207.
- Harley, T.A. (2014). *The psychology of language: From data to theory*. London: Psychology Press.
- Huey, E.B. (1908). *The psychology and pedagogy of reading*. New York: Palgrave Macmillan.
- Jacobs, A., Schuster, S., Xue, S. and Lüdtke, J. (2017). What’s in the brain that ink may character: A quantitative narrative analysis of Shakespeare’s 157 sonnets for use in neuro-cognitive poetics. *Scientific Study of Literature* 7(1): 4–51.
- Jacobs, A.M. (2016). The scientific study of literary experience. *Scientific Study of Literature* 5(2): 139–170.
- Kirwan, P. (2015). *Shakespeare and the idea of apocrypha: Negotiating the boundaries of the dramatic canon*. Cambridge: Cambridge University Press.
- Koppel, M., Schler, L. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the Association for Information Science and Technology* 60(1): 9–26.
- Kuiken, D. (2016). The implicit erasure of “literary experience” in empirical studies of literature. *Scientific Study of Literature* 5(2): 171–177.
- Kulynych, J. (2002). Legal and ethical issues in neuroimaging research: Human subjects protection, medical privacy, and the public communication of research results. *Brain and Cognition* 50(3): 345–357.
- Laeng, B., Sirois, S. and Gredebäck, G. (2012). Pupillometry a window to the preconscious? *Perspectives on Psychological Science* 7(1): 18–27.
- Levine, S. and Horton, W.S. (2013). Using affective appraisal to help readers construct literary interpretations. *Scientific Study of Literature* 3: 105–136.
- Louwerse, M.M., Benesh, N. and Zhang, B. (2008). Computationally discriminating literary from non-literary texts. In S. Zyngier, M. Bortolussi, A. Chesnokova and J. Auracher (eds.), *Directions in empirical literary studies*. Amsterdam, The Netherlands: John Benjamins, pp. 175–192.
- Mahlberg, M. (2013). *Corpus stylistics and Dickens’s fiction*. London and New York: Routledge University Press.
- McCarthy, K. (2015). Reading beyond the lines: A critical overview of cognitive approaches to literary interpretation and comprehension. *Scientific Study of Literature* 5(1): 99–128.
- McGann, J.J. (1991). *The textual condition*. Princeton, NJ: Princeton University Press.
- McKenzie, D.F. (1986). *The Panizzi lectures 1985: Bibliography and the sociology of texts*. London: The British Library.
- O’Dair, S. (2014). All’s well that ends Orwell. In C. Carson and P. Kirwan (eds.), *Shakespeare and the digital world*. Cambridge: Cambridge University Press, pp. 115–127.
- Parente, F., Conklin, K., Guy, J., Carrol, G. and Scott, R. (2020). Reader expertise and the significance of small-scale textual features in prose fiction. *Scientific Study of Literature* 9(1): 3–33.
- Peskin, J. (2007). Constructing meaning when reading poetry: An expert-novice study. *Cognition and Instruction* 16: 135–263.
- Pickering, M., Frisson, S., McElree, B. and Traxler, M. (2004). Eye movements and semantic composition. In M. Carreiras and C. Clifton (eds.), *On-line study of sentence comprehension: Eye tracking, ERPs and beyond*. New York, NY: Psychology Press, pp. 33–50.

- Rudman, J. (2016). Non-traditional authorship attribution studies of William Shakespeare's canon: Some caveats. *Journal of Early Modern Studies* 5: 307–328.
- Shillingsburg, P. (1996). *Scholarly editing in the computer age*, 3rd ed. Ann Arbor: University of Michigan Press.
- Shingler, K. (2011). Perceiving text and image in Apollinaire's calligrammes. *Paragraph* 34: 66–85.
- Simpson, P. (2004). *Stylistics: A resource book for students*. London: Routledge.
- Sotirova, V. (2014). Production and intentionality. In P. Stockwell and S. Whiteley (eds.), *The Cambridge handbook of stylistics*. Cambridge: Cambridge University Press, pp. 132–148.
- Stockwell, P. (2005). Stylistics: Language and literature. In B. Aarts and A. McMahon (eds.), *The handbook of English linguistics*. Oxford: Blackwell, pp. 742–758.
- Stockwell, P. (2016). The texture of authorial intention. In J. Gavins and E. Lahey (eds.), *World building: Discourse in the mind*. London: Bloomsbury, pp. 147–163.
- Stockwell, P. and Mahlberg, M. (2015). Mind-modelling with corpus linguistics in David Copperfield. *Language and Literature* 24(2): 129–147.
- van Boxtel, A. (2010). Facial EMG as a tool for inferring affective states. In *Proceedings of measuring behavior*. Wageningen, The Netherlands: Noldus Information Technology, pp. 104–108.
- Vipond, D. and Hunt, R.A. (1984). Point driven understanding: Pragmatic and cognitive dimensions of literary reading. *Poetics* 13: 261–277.
- Willems, R.M. and Jacobs, A.M. (2016). Caring about Dostoyevsky: The untapped potential of studying literature. *Trends in Cognitive Sciences* 20(4): 243–245.
- Zeitz, C.M. (1994). Expert-novice differences in memory, abstraction, and reasoning in the domain of literature. *Cognition and Instruction* 4: 277–312.
- Zwaan, R.A. (1993). *Aspects of literary comprehension: A cognitive approach*, vol. 29. Amsterdam: John Benjamins Publishing.

English language and digital health humanities

Brian Brown

Introduction

In this chapter we will examine the newly emerging health humanities and the opportunities these afford for explorations of patterns of language and culture, as well as novel examinations of the context and culture of health and illness through the medium of language. The health humanities represent a new departure in scholarship and practice as they attempt to include the full range of health professions, as well as informal carers, creative therapies and service users themselves. In this respect they are deliberately broader in scope than the medical humanities and involve an ambition to incorporate occupations and constituencies which have so far been barely touched by the established medical humanities.

Our efforts to broaden the inclusivity of the medical humanities in this way has coincided with a growing recognition of the potential of information and communication technology in the humanities themselves. This has taken a variety of forms, but the term ‘digital humanities’ has emerged as a consensus concept to embrace the diverse and multifaceted movement to exploit technology within the humanities. Koltay (2016: 781) defines the digital humanities as ‘an interdisciplinary field of study, research, teaching and invention concerned with the intersection of computing and the disciplines of the humanities’. To put it another way, Svensson (2010) writes of ‘the study of what happens at the intersection of computing tools with cultural artefacts of all kinds’. So far, a good deal of what takes place in the digital humanities has taken the form of facilitating electronic access to the resources of the humanities scholar. These have included projects such as the development of digital intellectual resources to study and reconfigure the concepts we use to understand the human condition (Burman et al. 2015), work on digitising, indexing and disseminating historical materials (Royles 2016), as well as exploration of the potential of ‘big data’ to enable us to map the history of ideas (Flis et al. 2016).

Despite this flurry of activity to embrace the digital revolution within the humanities, there have so far been relatively few efforts to think through the relationship between the medical or health humanities and digital opportunities. Using the terms together in a search of major databases such as Scopus at the time of writing in 2017 yields a mere 7 entries – compared to nearly 1500 for digital humanities as a whole. Even fewer of these are directly

relevant to the purposes of this chapter. Therefore, writing it has been a difficult task because there are so far few maps or signposts as to what a digital health humanities might look like. At the end of this chapter we will turn to the question of how digital approaches in the health humanities can yield future directions to the digital humanities as a whole.

As a result of this, the drift of this chapter will be somewhat autobiographical. The author (Brown) has had the good fortune to work with some interesting and inspiring colleagues over the past decade to craft the field of health humanities and make efforts to utilise communication and information technologies to ensure its place in the academic and practitioner landscape. Significant in this enterprise has been the work of Paul Crawford, professor of health humanities at the University of Nottingham, and Charley Baker, also at Nottingham, whose work combines a deep and compassionate understanding of mental health nursing and a profound knowledge of speculative fiction and countercultural creative activity. Of course, many others have contributed to the health humanities enterprise, but it was with these two individuals that I began a project to look at representations of madness in English language fiction in 2007. We were fortunate to obtain a grant from the Leverhulme Trust to support this, and the project resulted in our book *Madness in Post-1945 British and American Fiction* (Baker et al. 2010). As we argued at the time, the last couple of millennia of human storytelling are full of accounts of bizarre or inexplicable experiences, mental distress, behavioural disturbances and interpersonal difficulties. Fiction has also been at the forefront of depicting fantasy, resistance, resilience, tenacity, resourcefulness and creativity which can all, depending on context and circumstance, be labelled as deviant. While the nineteenth century had been relatively well discussed (e.g. Showalter 1985) there was at the time very little critical scholarship on the late twentieth century or the contemporary era. Bringing these stories together in a critical context was valuable for scholars and practitioners. Like much scholarship concerned with the humanities in health, it helps extend empathic reach (Charon 2017; Parker 2018) and as participants in our health humanities conference series have told us, provides a valuable map of unusual experiences, distress and disorientation. The humanities have a vital role in understanding, ameliorating and offering succour in health and illness, and indeed are increasingly being used in the education of doctors and other healthcare professionals (Graham et al. 2016; Greene and Jones 2017). As we have documented elsewhere (Crawford et al. 2014), departments and teaching programmes have sprung up in universities around the world under the heading of ‘medical humanities’. In our ‘madness’ writings, the question of whether the depictions of mental ill health are ‘good’ or ‘bad’ compared with clinical accounts is not the major one. After all, very few everyday examples of so-called mental disorder match exactly the descriptions found in introductory textbooks exactly. What was more important in our view was to show the diversity of experiences that could be formulated and embedded in narratives and the value of these in making sense of the human condition.

Background

Up to this point the journey had been interesting and challenging, but it was not specifically ‘digital’, despite the extensive use of computers to manage databases, texts and communication. The breakthrough came with a small Arts and Humanities Research Council grant to create a Madness and Literature Network (www.madnessandliterature.org) which is still live to this day. Here, visitors can create accounts, examine others’ reactions to books and contribute their own. It has also served as a platform for promoting conferences and other meetings. It provides a resource that we lacked when first embarking on our madness and literature project in the first place – a bibliography of literary resources which represented

madness. Of course, most people can think of a few, but the books dealing with this theme are legion, and no one had gathered them together on one place before. This has come about through the power of what we might call, in the contemporary idiom, crowdsourcing. In the wider digital humanities literature there has been considerable interest in crowdsourcing (Carletti 2016; Edelstein 2016). The digital humanities offer opportunities for lay and public participation in the creation of knowledge and culture in a way that is fundamentally different from the earlier paper-based versions. Carletti (2016) documents how in recent years, a number of projects have been initiated by cultural and educational organisations to exploit the capacity of information and communication technology to build novel resources relating to heritage and history. This involves exploring new forms of partnership with the public. Carletti notes that among the emerging models of collaboration, a process she calls ‘heritage crowdsourcing’ has been widely used to construct novel histories. It involves the online involvement of the public (the “crowd”) in providing data, information and knowledge as a result of institutional calls for contributions.

Carletti (2016) notes that these kinds of collaborations between the academy and the ‘crowd’ have served to increase engagement with digital literary heritage and co-create digital content. Indeed, technological advances are breaking down barriers between experts and laypeople. In some cases, misgivings have been expressed about crowdsourcing as a way of generating content, both in terms of the intellectual quality and in terms of the way it often involves people giving their labour free. Even so, it represents an increasingly important way of facilitating access to content and of content generation itself in fields as diverse as literature, heritage, museums and science. As Carletti notes, projects may vary in terms of what they ask the crowd to do. In some cases, as with citizen science, the crowd answers a limited range of pre-defined questions set by the project’s originators. In other cases, there may be scope for a good deal more public involvement in collaborative and co-created activity, where the public or community contributors have a good deal more creative control over the scope and direction of the project.

Critical issues and topics

Our involvement in the madness and literature project prompted some further reflection on the relationship between humanities and health. Of course, a variety of activities had been going on under the rubric of the medical humanities for some time, but in 2008 we became increasingly aware that more could be done with a greater variety of practitioner and client groups and a wider panoply of humanities and creative disciplines than had hitherto been seen. In other words, we became aware of a growing need for a new kind of debate at the intersection of the humanities and healthcare. In the recent past the field of medical humanities had grown rapidly, but in the second decade of the twenty-first century it was timely and appropriate to address the increasing and broadening demand from other professions to become involved and to accommodate new sectors of the healthcare workforce and the public. There were important cohorts of personnel in healthcare, a whole army of ancillary workers, as well as informal carers and patients themselves who had been largely left out of the medical humanities up to that point. Moreover, as different disciplines came to value the contribution made by the arts and humanities and new opportunities emerged in health for the development and inclusion of new approaches in the humanities, it is important that this expansion and debate was given voice and new forums were created for these new developments. This is why we started talking instead of the health humanities to indicate this broader focus and wider range of stakeholders.

The medical humanities were, of course, the first on the scene, and have developed strongly in the Anglophone world. But now the field of health humanities is fast developing a more inclusive and international capture of material as other disciplines and different nations develop their own distinctive practice and theory. Its expansion is marked by health humanities-focused research funding calls, dedicated centres and networks and changes of existing medical humanities centres to align with health humanities, either by adopting more inclusive names and mission statements or by undertaking new programmes of work. With increased consideration of the emergence of health humanities in the UK, United States and Canada and resultant new inflections of medical humanities as ‘critical medical humanities’, as in Bates et al. (2014), all this merited a new publication aiming to give a flavour of the full range of healthcare activities and the newly discovered relationships between these and the humanities themselves. Hence our 2014 book *Health Humanities* (Crawford et al. 2014). Indeed, the rise of the more inclusive health humanities marks an evolution of medical humanities.

In some cases, this growth of interest from scholars and practitioners in the humanities in healthcare has been spearheaded by the medical humanities themselves. After all, this field has now become a ‘mature discipline’ (Ahlzen 2007). Certainly, this area has been influential and has recently gained ground as an alternative to the traditional, exclusively scientific curriculum pursued in the English-speaking world and many postcolonial nations. Changes in medical education were prompted by re-evaluations from educators and professional bodies on both sides of the Atlantic through the 1990s (Christakis 1995; General Medical Council 1993; Enarson and Burg 1992; Schwarz and Wojtczak 2002). This has involved a great deal of questioning of medical school curricula and a reappraisal of the kinds of qualities that educators are attempting to foster in tomorrow’s health professionals. There has been some reflection on the role of digital technologies in the delivery of medical humanities curricula. Kemp and Day (2014) highlight the value of medical humanities components in medical training in fostering a tolerance of ambiguity, enabling creativity, enhancing understanding of patients’ experience of illness, enhancing reflecting learning and enabling learners to appreciate the viewpoints of others. They consider using digital technology to challenge, change or add to existing beliefs or knowledge. There are a whole variety of well-known software and web-based means of delivering curriculum content. For example, the Blackboard system is widely used in UK higher education, and some universities have made learning materials accessible outside the institution, such as Harvard University’s massive open online courses (MOOCs), as well as a proliferation of stand-alone websites and YouTube content. In some of these systems there are possibilities of recipient tailoring by adapting the release of materials to the learners’ pace, style and learning needs. Yet the challenges of learning in the new millennium go deeper than this. There is a desire to make collaboration an important part of learning in medical humanities programmes (Bleakley and Marshall 2013). Here, digital technology can be of assistance, in facilitating collaboration, capturing creative output, facilitating digital storytelling and, as in the University of Michigan, virtual exhibitions of student projects can be held (Kemp and Day 2014).

The development of medical humanities then has led to the expansion of curricula for healthcare professions to include the humanities. Classically, in the case of medicine, this has involved ethics or ‘moral attitude’ (Olthuis and Dekkers 2003), or confronting students with some of the enigmas or conundrums of the discipline to make them aware of philosophical issues (Brawer 2006). More recently this has broadened to include literature (Dysart-Gale 2008), expanding clinical empathy (Garden 2008), dealing with the more exasperating experiences of clinical life (Garden 2008), as well as developing community education and

a commitment to interdisciplinarity (Donohoe and Danielson 2004). In addition, we have seen a focus on both medicine and the humanities as interpretive enterprises (Gillis 2008), or recognition of healthcare practice as a kind of performance, analogous to being a musician (Wooliscroft and Phillips 2003). Within the medical humanities, one of the corollaries of this is that medicine has reconfigured its boundaries to become interdisciplinary. At the same time, it has become disciplined through the humanities on the premise that ‘arts and humanities approaches can foster significant interpretive enquiry into illness, disability, suffering, and care’ (Bolton 2008: 131). The notion of ‘humanities’ itself suggests a shared understanding of what it means to be ‘humane’, either as a person or as interpersonal practice. Some of the purported benefits of teaching medical humanities include the promotion of ‘a patient-centred approach to medical care’, ‘counteracting professional burnout’ and ‘equipping doctors to meet moral challenges not ‘covered’ by biomedicine’ (Gordon 2005; Petersen et al. 2008: 2). Yet this diversity of claimed benefits is often as confusing as it is comprehensive. Thus, there are those who point to the lack of consensus as to what exactly constitutes medical humanities and what the discipline is for (Petersen et al. 2008).

Until recently then, there has been an apparent hegemony of medicine in the field. Our insistence on the alternative term ‘health humanities’ has grown from a belief that medicine does not have exclusive claim upon the action. As we outlined in the *Health Humanities* volume, a variety of other disciplines have sought to incorporate the arts and humanities, or have attempted to relate the arts and humanities to practitioner education, practice with clients, professional development and the wider development of the discipline. The hitherto more narrowly defined medical humanities therefore do not necessarily have a monopoly over the work undertaken. In particular, the vast body of practical work going on around the world in the expressive therapies demands and deserves admission to the healthcare humanities. This includes but is not limited to art therapy, music therapy, dance therapy, poetry therapy and related therapeutic disciplines that involve some sort of creative and expressive activity. Here, practitioners and researchers are often actively seeking places where they can develop ideas and place them on a more theoretically mature footing, as well as providing a means of bringing this work more fully to the attention of scholars in the medical humanities, those in practice and those who teach the health professionals of the future.

The humanities can address issues which are common to people in a variety of healthcare roles. In today’s ‘fast healthcare’ (Crawford and Brown 2011) clinicians from all backgrounds can feel under pressure to adopt a ‘production-line mentality’ and lose sight of the human in the process. Fictional narratives can offer a restorative ground for clinical being – a way to think about and feel for the person affected by, for example, psychosis, depression or dementia. Novels, short stories, memoir, plays and poetry can help us develop what Paul Gilbert calls a ‘compassionate mind’ (Gilbert 2009) and consider the rich interplay between social, political and cultural perspectives on health and illness.

Current contributions and research

To what extent does the digital revolution in scholarship offer new opportunities for the humanities in health? As we have remarked earlier, the expansion of digital technologies and the augmented role of the humanities in health have facilitated shifts in the territories of expertise and patterns of participation found in health and social care. The generation of content, discussion of problems, airing of grievances and complaints as well as sharing good practice has all been speeded up. There are a variety of well-known Internet forums where people can upload and share experiences of health and illness. For example, the accounts available

at www.healthtalk.org/ have been used for educational and research purposes and for people unlikely to receive ethical approval for first-hand research with patients (such as students) these kinds of facilities offer unrivalled opportunities to engage with the experience. In addition, University of Nottingham scholars associated with the Health Language Research Group have compiled a corpus of language concerning health from a variety of studies. This is written or transcribed spoken language and comprises data from a study of telephone helpline talk, health language in online settings and a number of healthcare encounters. As a result of an ESRC funded study the University of Lancaster has compiled a corpus 40 million words of feedback about the UK's National Health Service (NHS). Furthermore, online discussion boards, video sharing sites and other social media are becoming resources for researchers.

As well as web-based resources specifically geared toward health and illness themes, these more general facilities which users have employed to explore health and illness are rich sources for researchers. For example, the question of people making videos about mental health and posting them on video sharing sites such as YouTube, is described by Fletcher (2016). Fletcher considers the way in which people present video diaries, or 'vlogs', of their experience with schizophrenia. This has the potential not just to represent experience but to change it. In speaking in video excerpts about schizophrenia, in combination with comments generated about them by others, users can represent transitional, dialogical states of illness and create resistance against static definitions of schizophrenia. In doing this, users might foster creativity, experimentation and inventiveness so as to transform understanding of schizophrenia within the sciences and humanities, thus making it possible through visual media and the interaction it promotes to push back at the definitions of schizophrenia that science and medicine have created.

The expansion of digital humanities has also resulted in some shifts in professional and institutional life. This has included changes in the role of professions such as new opportunities are opening up for librarians and information professionals as the compilation and curation of resources becomes more significant (Burns 2016). Indeed, the burgeoning literature on digital, humanities itself contains a major focus on matters of curation and librarianship. This creation of new content and new ways of accessing it has led some authors, such as Clough (2016) to propose that the humanities and the social sciences are involved in a new regime of calculation, quantification and measure, the opportunities for which are facilitated by the expansion of electronically held and computer mediated information. Clough refers to this as the 'datalogical turn' to denote 'the data mining of social media, tracking devices, and biometric and environmental passive microsenors – the full analytic capacities of twenty-first-century media – now deployed not only in but also outside the academy in business, culture, government, economy, and individuals' lives' (Clough 2016: 437).

Health humanities, then, are facilitating and are afforded new opportunities by the rise of big data and associated attempts toward greater surveillance and control as data are collected by governments, political and pressure groups, security and police services and Internet utilities such as Google, Facebook and Twitter, as well as mobile telephony providers, to name only the most usual of the usual suspects. It is not just a matter of harvesting data – the data have to be organised, indexed and made sense of too, for them to be meaningful for other human and political actors. This does not necessarily always have authoritarian consequences. We are often made aware of concerns about the surveillance society and the database state (Acquisti et al. 2015; van Heek et al. 2017), but the pursuit and excavation of large datasets can have liberal, if not emancipatory, consequences too. For example, the recent ESRC-funded study of feedback on the NHS by Paul Baker and Gavin Brookes has brought to light a variety of everyday dissatisfactions, for example, relating to

appointment availability, the difficulty in dealing with receptionists and the feeling that staff are not taking patient concerns seriously. While we only have preliminary reports so far (www.lancaster.ac.uk/news/articles/2017/study-finds-politeness-compassion-and-a-sense-of-humour-prompt-nhs-feedback/), this kind of work can help tilt the balance in favour of the consumer or patient in a field often dominated by policy or professional voices. We will explore the emancipatory possibilities of looking at large corpora of language in discussing our own work on discourse around mental health later.

Sample analysis

Looking at digitised language: two examples

To illustrate this let us consider two examples from our own work. The first of these integrates concepts from the humanities in the form of literary theory with linguistics and with more conventional social science topics such as the representation of mental health in the media. This has been undertaken as part of our AHRC-funded study of recovery in mental health. The project as a whole is entitled *Creative Practice as Mutual Recovery* and comprises a number of work packages addressing activities such as sculpture, music, storytelling photography and yoga and their role in facilitating improvements in wellbeing and social connectedness. In addition, there were sub-projects concerned with textual analysis, including mental health in the media and in policy discourse which aimed to understand how mental health and arts participation were framed and how relationships between arts participants were understood. The concept that we have adopted to study this is that of ‘voice’.

The idea of looking at news texts in terms of ‘voice’ takes its cue from literary theorist Mikhail Bakhtin (1981) who in the early twentieth century emphasised the way that many texts are formed from a variety of voices. In this particular case, analysing mental health in the news, we may not often think of news articles as polyphonic, that is involving many ‘voices’. However, as Montgomery argues, in most news ‘the discourse supplied by the institutional voice is supplemented and elaborated by a range of other voices – some, such as correspondents and reporters, affiliated to the institution, some, such as politicians, bystanders, experts and witnesses, not’ (Montgomery 2007: 87–88). In linguistics, this approach has been promoted by Bednarek (2016) who examines voice in news media in terms of who is speaking, where the voices are located, how the voices are identified and how other voices are integrated into the narrative. When writing news articles, journalists adopt a position and create a voice, which can be corroborated or contested by including additional voices (Calsamiglia and López Ferrero 2003). Such polyphony in utterances is typically examined from the vantage point of discourse analysis and linguistic analysis frameworks (Fløttum 2010; Fløttum and Stenvoll 2009; Gjerstad 2007; Wodak 2009). In these analyses, direct quotation has been highlighted as the most explicit form of incorporating others’ voices into one’s own and a mechanism for ‘managing the words of others to convey and serve the purpose of the writer’ (Calsamiglia and López Ferrero 2003: 149). Questions like who is quoted, how often and how they are quoted, for example, directly or via indirect quotes (Bednarek 2016; Calsamiglia and López Ferrero 2003). Their ‘voice’ is also shaped by how they are introduced in pre-quotation segments (van Leeuwen 1996) and what they say. Coverdale and Nairn (2006) have, for example, suggested that if mentally ill individuals were more often directly quoted, media representations of mental illness would have been more positive.

Our examination of these issues was undertaken via an analysis of 1412 newspaper articles published in the UK national and local press between 2007 and 2015. The year 2007

was chosen as a starting point because in this year a new Mental Health Act was introduced and 2015 was the last full year up to the point of study. The Nexis database of news articles (www.nexis.com/) was searched for material containing terms such as ‘mental health’, ‘mental illness’ and ‘art’. Of the items in the final sample 339 national newspaper articles were included as well as 1073 local press items; the latter were of particular interest, as it is in these that initiatives for mental health service users were most likely to be reported. The study itself is described in more detail in Atanasova et al. (2019). With the 485,000 words of material acquired from the database, our analysis began by identifying those words which appeared significantly more frequently in the news than in the English language as a whole with the help of Wordsmith Tools (Scott 2016), using the British National Corpus (2009) as a comparison.

This provided a means of identifying so-called keywords, which, according to McCarthy and Handford, are words which ‘best define’ (McCarthy and Handford 2004: 174) a text or texts – words that are significantly higher in one language variety compared with another. Keywords are an important indicator of both expression and content (Seale et al. 2007) and have been used by an increasing number of researchers as a reliable means of identifying key themes in characterising health language corpora (Adolphs et al. 2004; Seale 2006; Harvey et al. 2007; Seale et al. 2007). Identifying keywords has been a basic staple of much corpus-based analysis, for example, in areas as diverse as examining knowledge claims in scientific discourse (Plappert 2017), examining emotions in language (Jain et al. 2017) and lyrics in popular music (Motschenbacher 2016). In this sense, a ‘keyword’ is one which is statistically more frequent, rather than the usual sense of words deemed to be of significant social and cultural importance.

In this analysis, frequently occurring words (or ‘keywords’) were grouped into semantic sets, which included ‘activism’ (keywords including AWARENESS [328 occurrences], RAISE [448 occurrences]); ‘art’ (e.g. MUSIC [353 occurrences], EXHIBITION [383 occurrences]); ‘institutions and infrastructure’ (e.g. WORKSHOP [277 occurrences], COUNCIL [265 occurrences]); ‘medicine’ (e.g. DISORDER [590 occurrences], PATIENT [924 occurrences]); ‘relationships’ (e.g. COMMUNITY [750 occurrences], GROUP [1204 occurrences]).

As our interpretation of the data evolved, the insights generated by this process were regrouped into frames which concerned recovery, stigma and economic issues. Drilling down into the detail of news items relating to these frames was revealing. In the case of ‘recovery’ for example there were many items that quoted clients of services directly such as ‘former mental health ward inpatients’ saying ‘[m]usic is my passion and I believe it definitely enhanced my recovery’ (*Leicester Mercury* 18/09/14) and service users discussing how they could ‘conquer depression through the power of art’ (*Derbyshire Times* 15/02/13). By recounting success stories of how ‘pursuing artistic skills’ allowed sufferers from ‘a grey cloud of depression’ to ‘add colour back’ into their lives (*East Anglian Daily Times* 06/10/12) and of individuals successfully ‘combating’ mental illness through ‘the healing power of creativity’ (*Liverpool Echo* 16/10/14). Where the question of stigma was concerned, this pattern of quoting service users was also prominent. For example, where arts initiatives were presented as a means ‘to challenge the stigma around mental health issues’ (*The Plymouth Herald* 05/04/12) and more generally raise awareness. Stigma was characterised in terms of ‘[i]gnorance and superstition’ (*The Guardian* 23/12/14); and the combined effect of ‘misinformation’ attached to mental illness (*Bath Chronicle*, 11/07/13) and lack of information – ‘people don’t know any better. If no one around them is talking about mental health, and it’s something that’s “hushed up” all the time, where are they going to learn’ (*The Western Mail* 24/02/14).

Thus, the inclusion of others' voices can be a means of conveying the veracity and immediacy of news and be a mechanism for constructing newsworthiness. Voice and news values are related and Bednarek (2016) has recently argued that bringing together the analysis of news values and the analysis of voice can provide more detailed insights into journalistic practice. In news reports about natural disasters, for example, newsworthiness is often established through the accounts of eyewitnesses describing the extent of the damage (suggesting news values such as 'bad news' and 'magnitude') and how surprising the event was for them (invoking the news value of 'unexpectedness') (Bednarek 2016). A similar phenomenon is described by Wooffitt (1992) in his phenomenon of 'active voicing' – the use of quoted speech to establish the factuality of a claim. Rather than simply the writer, the active voice brings into being new corroborating actors, who have opinions and personalities of their own. Analysis of what they say, what is being said about them and by whom, how their voices are included – direct quote, paraphrase, partial quote or unattributed (Bednarek 2016) and how they are introduced (van Leeuwen 1996) can lead to a better understanding of mental health in the media. This need to investigate whose voices speak through news reports is heightened by journalists' ever tighter deadlines and shrinking resources – factors that can be expected to increase the influence of readily accessible sources on news content (Schudson 2003).

While the investigation we have described has been concerned with matters of representation and voicing and is driven by techniques in literary study and linguistics, it has nevertheless some important implications. Our finding that service users in particular are represented as viewing creative practice as therapy has important implications. Doing the artwork or singing may have a therapeutic effect on the troubled individual and acknowledging this is a significant and encouraging trend from the point of view of those seeking to promote good news coverage about mental health. What may be equally significant, but which is less apparent in press coverage is that these kinds of activities enhance social relationships and may serve to make relations between service users, informal carers and professionals more egalitarian and mutual (Brown 2015). Moreover, the activities themselves change the institutions in which they take place – a mental health facility with a thriving art therapy programme is a rather different place than one without.

Yet these broader social and institutional aspects are not so evident in the analysed coverage. The simplistic representations of art as therapy that we found feed into concerns that art and creative practices may be becoming 'anaesthetic interventions' (Mirza 2006). The way initiatives of this kind are represented emphasise personal struggles and journeys of healing, as has been observed by other authors seeking to understand the social representation of mental health (Brown and Baker 2012; Olstead 2002). The creative activity is posed as a kind of therapeutic or 'emollient' process facilitating recovery. The intervention is something that the service users consume to enable the management or diminution of symptoms. As Cochrane put it in 1989, such interventions tend to 'clientise' people who are already likely to be disempowered and impoverished. Without disregarding the benefits and often the sheer enjoyment of engaging in creative activity, there is often limited focus on the sociopolitical dimensions of empowering people. They do not necessarily gain material powers or become significant actors in the design and delivery of services (Cochrane 1989; Brown and Baker 2012).

A distinctive feature of the data described here was the co-occurrence of *recovery* and *stigma* frames in news coverage, which appeared together on over two-thirds (68 percent) of the occasions where they were identified. While many previous accounts of media coverage of mental health issues have highlighted stigma and related this to inopportune media

representations (Klin and Lemish 2008; McGinty et al. 2016), here stigma was discussed as an undesirable phenomenon that needed to be combatted. The frequent occurrence of the *recovery* and *stigma* frames together, reflects the Department of Health implementation strategy on mental health and recovery, which argues that '[t]ackling stigma and discrimination and promoting human rights are at the heart of this strategy' (HM Government 2011: 29). What is noticeable about the press coverage we studied, as well as much policy and professional discourse (Baker et al. 2013), is that stigma is always seen as coming from ordinary people or society as a whole. Professional and policy-making actors are not identified as sources of stigma. This is despite many academic writers drawing attention to the persistent negative representations of people with mental health problems as risky, vulnerable and having limited potential in professional and policy discourse (e.g. Chamberlain 2016). Moreover, there are reports of clients suffering discriminatory treatment in therapeutic programmes themselves (Holley et al. 2016). Yet news media representations identify stigma as emerging from wider society rather than policy and practitioner groups.

To sum up then, the approach outlined in our first example, combining features from literary and linguistic study can illuminate to hitherto unstudied and underappreciated aspects of media coverage, such as the role of 'voice' in the news and the absences as well as the presences which can tell us about the sociopolitical situation in which the texts operate and perhaps also about the lives of the people involved.

The distinctiveness of our approach was underscored for us by the literature we reviewed in the process of undertaking this study. We found that published studies on the mass media and mental health mostly employed a deductive content analysis approach (e.g. Aragonès et al. 2014; Bilić and Georgaca 2007; Everett 2015; Houston et al. 2016). Despite the apparent objectivity of this approach it is lacking in flexibility and may lead to the omission of new themes. Studies using qualitative techniques in which themes are derived directly from the analysed media content are, in contrast, flexible, but more dependent on researchers' interpretations and therefore, seen as less objective.

In addition, our work on mental health in the mass media highlighted for us how little work there was on representations of this phenomenon in social media. While social media was generally understudied, we also found that most existing research focused on testing the utility of social media analysis for the identification of at-risk individuals. Such use of social media is, indeed, a rapidly growing area of research variously referred to as info-veillance or digital disease detection (Conway and O'Connor 2016). In contrast, research into the content and potential impact of social media messages on opinions about mental health, which can provide valuable insights about public opinion and policy support, was rarer. More research on the social media messages related to mental health and their potential impact is needed, not the least because traditional media's 'singular power' in influencing issue interpretation may be weakening in our networked society (Meraz 2011: 107). The need for more such research is also underlined by evidence that people experiencing mental health problems use social media 'as part of their recovery, to live well and to challenge stigma' (Betton and Tomlinson 2013: 215).

To illustrate the potential of some of the approaches we have outlined in examining the content of Internet material, let us take a second example from our own work. This is not exactly 'social media' per se, but is interesting because it shows the potential of the techniques we have adopted in examining the large amounts of text generated in Internet contexts. The approach we have taken is, like the media study of mental health described earlier, based on the numerical analysis of larger bodies of language. Such collections of text, or 'corpora', can be systematically explored for the insight it can give about how a language

is used, terms are deployed, what forms and terms tend to occur together and the concepts into which it can be formulated. Compilers of dictionaries have used this for some time to learn what words are used and what they mean in a particular language community, and this approach has been gaining ground in second language learning and teaching (Boulton and Cobb 2017; Lee et al. 2017). Over the last few years we have been applying this technique to the analysis of healthcare language. We collaborated with the authors of a former website set up to offer health advice to young people, www.teenagehealthfreak.org.uk which, among various other features, offers an online agony aunt service. Over the period 2004–2008 we were able to interrogate a 1.6-million-word corpus of questions written to the website by its users. The data therefore provide a substantial snapshot of health concerns communicated on a regular basis by the contributors. The examples and analysis we will consider here are discussed in more detail in Harvey and Brown (2012) and Crawford et al. (2014).

To begin interrogating this substantial body of language, we generated a list of keywords which provide a thematic characterisation of the language. In our case, the keywords were operationally defined as those which appeared significantly more frequently than expected by comparison with the British National Corpus. Table 27.1 presents a representative overview of these keywords resolved into health-related themes.

As can be seen from Table 27.1, the most frequently occurring terms can be grouped into meaningful clusters of content items indicative of the central health issues which concern the website users. Although informed by clearly defined semantic areas, the categories overlap to some extent. For example, mental health, body image, drugs and alcohol all have links to mental health. Clearly, reproductive health yielded a great many queries. We have discussed these at greater length elsewhere (Harvey et al. 2007). Here, let us follow Harvey and Brown (2012) and consider what the *Teenage Health Freak* corpus can tell us about self-injury. Given the prevalence of the phenomenon in the contemporary era, with as many as 47 percent of adolescents affected in some studies (Cipriano et al. 2017), it is not surprising

Table 27.1 Keywords by categorisation of health themes in adolescent health emails

Theme	Examples of keywords
Sexual health	Sex, sexual, penis, pregnant, period, orgasm, AIDS, infertile, STD, STI, sperm, contraception, HIV, clitoris, vagina, vulva, PMS, erection, condom, masturbate, gay, abortion, foreplay, intercourse, virgin, unprotected, lesbian, oral, pill, ovulation, herpes, thrush, chlamydia, pregnancy, tampon, testicles, genitalia, viagra, scrotum, labia, glans, ovaries, foreskin, balls, fanny, bisexual, miscarriage
Mental health	depression, depressed, suicide, suicidal, die, overdose, antidepressants, cut, cuts, cutting, self-harm, scars, prozac, sad, unhappy, self, harm, wrists, addiction, addict, stress, stressed, ADHD, paranoid, mental, mad, moods, sad, unhappy, crying, personality, anxiety
Body weight/image	Anorexia, anorexic, weight, size, overweight, fat, obese, underweight, skinny, thin, bulimia, BMI, exercise, diets, kilograms, KG, KGS, bulimia, calories
Drugs/alcohol	Drugs, cannabis, cocaine, heroin, pills, alcohol, drunk, drinking, poppers, mushrooms, marijuana, crack, ecstasy, addict, stoned, LSD, cigarettes, dope
Serious conditions	Cancer, epilepsy, diabetes, anthrax
Minor conditions	acne, zits, blackhead, mumps, scabies, dandruff, worms, cystitis
Medication	Medicine, medication, prescribed, antibiotics, tablets, pill, pills

Source: Harvey and Brown 2012: 322

that in a corpus of this size there were examples of people describing and soliciting help for their self-injury. Terms such as ‘self harm’ and closely related ones such as cutting and self harmer occurred 790 times in the corpus, considerably more than one would expect by chance, compared with the British National Corpus. In Table 27.2 we present some more data from Harvey and Brown (2012) where we have examined the occurrence of self-harm with terms such as ‘help’ and ‘stop’. This was done with Wordsmith Tools’s Mutual Information scores. We examined these terms because they appeared to offer some potential for insight into the dilemmas surrounding the desire to give up self-harm and adopt different coping mechanisms that do not involve self-injury. This exemplifies how the digital tools that are used to interrogate corpora allow the analyst to examine its use in context in order to better understand how and why the word is being used in a certain way.

The examples in Table 27.2 yield more detail concerning personal experiences of self-harm, with several new themes visible. For example, self-harm is said to be a response to physical and sexual abuse, verbal bullying, family turmoil and a reaction to feeling upset and depressed. Self-harm is constructed as a survival mechanism, a form of relief, or as one contributor put it, a means of trying ‘to deal with it’, in this instance, the trauma of sexual abuse. Self-harm also appears to serve as a visual manifestation of distress, although the actual physical injuries that speak for those who cut (Ross 1994: 13) can also be a source of disquiet: ‘I really want to stop the scars are really ugly!!’, ‘i hate having to hide my cuts/scars’. Thus, these accounts reveal the some of the intolerable circumstances of young people involved in self-harm. Although the act is portrayed as a survival mechanism, a response [no matter how ‘maladaptive’ to unbearable emotional distress (Favazza 1996)], its legacy of cuts and scars, which have to be treated and hidden from others, only appears to aggravate the predicament of the young people for whom self-harm is experienced as an effective form of relief. As one of the correspondents succinctly sums up their situation, ‘it’s a catch 22 situation Help!’

Table 27.2 Examples of messages containing the common collocates ‘help’ and ‘stop’

-
- i tried to **stop self harming**, and i managed 5 weeks 3 days, and i failed after that, and i know ive let myself down but alot has happened to me, which made me self harm in the first place, i kept things inside me for soo long, and i have suport from skool which is gr8 help, like i have my mentor and another teacher, but recently ive started to keep it inside me and is making me self harm, and i always wrote to my teacher to say how i felt, but something stops me now, and the alternatives i used instead of **self harming** dont **help** nomore. what can i do. i know i need help, but i cant stay strong with the self harming, so much hurt is inside me and i cant get it out of my head.. please please relpy!!
 - i need 2 **cut** myself all the time. **help** me pls
 - i have recently started **cutting** myself and i want to **stop** but i cant. ive treid loads of things and im woried becace when ever i get upset i start to visualise myslef cutting. and then when im alone if i havent calmed down i start **cutting** and i want to **stop**. i hate having to hide my **cuts**/scars and i need **help**
 - i smoke but im not addicted it just makes me better as i get easily stressed. have you got any suggestions for things i could do instead because i started smoking to **stop** myself from **cutting** myself
 - Doctor Ann: About a year ago I **stopd cutting** myself with a penknife because I’d managed to sort out my life. The other day my family had a row and I started to **cut** myself using kitchen knives, straight across my wrists. One of my friends saw the scars today, and everyone thinks I’m doing it for attention, but them thinking that partially lead to it anyway – it’s a catch 22 situation **Help!**
 - i self harm and i whant some **help** i hate my self for ding it it is like a vitiose circle is there any thing u can do that would help me to **stop** i do not whant to talk to my parents
-

Source: From Harvey and Brown 2012: 326

Interestingly, self-harm is also perceived to be an alternative to other risky behaviours. In an effort to avoid it, the adolescents seek participation in other deleterious activities, such as smoking. Yet such is the pull of cutting, these substitutes are comparatively less compelling and so abandoned in favour of self-harming behaviours. As one of the adolescent's comments: 'the alternatives i used instead of self harming dont help nomore' (sic). The pull of self-harm over and above these other risk behaviours can perhaps be further explained by the broad range of motives presented by the adolescents for the activity. As an antidote to psychological upheaval, self-harm is hard to replace since it can be caused for a number of reasons, while at the same time fulfilling different functions (Horne and Csipke 2009: 656), be it emotional release or articulating (such as visualising) inexpressible distress.

As well as representing a form of 'digital health humanities', using datasets such as we have described here can have a valuable contribution to make to the education of health professionals. A specialised corpus such as we have described has great potential for researchers and practitioners as a means of studying highly sensitive concerns from a generation who have often been reluctant to consult practitioners, peers and others for personal health advice and information face to face (Suzuki and Calzo 2004). The corpus approach to health communication affords an effective means of identifying the 'incremental effect' (Baker 2006: 13) of patterns across large quantities of text, allowing the researcher to discover linguistic patterns which are apt to remain submerged in extensive datasets. The material in corpora such as this might well yield important and unexpected opportunities for advancing healthcare communication, for both second language learners of English in healthcare work, where a data-driven approach is gaining momentum and in terms of a wider applicability in health communication learning as a whole.

Linguistics has not hitherto been central to the medical humanities or the newer health humanities. However, we have tried to show, by way of our work on press coverage of mental health and young people's problems written in to a health website, that it is possible to apply techniques from lexicography and second language learning such as corpus linguistics to elicit the patterns in larger datasets. These studies also highlight that there is language to explore in unexpected places which may provide insights into groups that have tended to be under-served by face-to-face healthcare.

Future directions

There are other examples that could be adduced, from using information technology to facilitate access to the arts in hospitals to studies of interaction between practitioners and clients, to studies of storytelling involving patients, carers and practitioners. However, I thought it more useful to describe a couple of examples (language in the media and language online) in some detail so the reader can more readily appreciate what can be seen in these kinds of investigations and how inferences can be drawn from them. In this chapter we have tended to focus on language and textual forms because that is where the majority of our investigations so far have lain. With the development of this approach there are some intriguing possibilities emerging from other media of expression such as music and the visual and three-dimensional arts. As we have argued elsewhere (Brown et al. 2006; Crawford et al. 1998), the use of language is central to healthcare practice and everything – from the whole life narrative to the minute details of conversational interaction – represent data from which a great deal can be gleaned for research and education.

As sociologists and anthropologists of medicine have argued, talking about one's health, or that of other people is a very widespread activity, throughout the majority of human cultures (Kleinman 1988; Bury 2001). Now that the digital sphere embraces so much of this discourse, the digital humanities can open up a variety of new ways of making sense of it. This material does not necessarily give us access directly to processes of sense making, consciousness or mindedness on the part of the originator. Nor does it provide direct knowledge of the original social context in which the story took place or within which it was read and understood. What is indisputable, however, is the role of material such as we have described in enriching the picture of health and illness; its potential to enhance understanding between practitioners, carers and clients; its role in educating practitioners; and its ability to build solidarity between sufferers themselves. The burgeoning field of health humanities is a testament to the desire to know something more than can be assessed simply through symptom checklists or investigation of biomarkers. The health humanities 'are not just a nice side dish of true medicine, but they are an indispensable part of patient oriented medicine' (Widder 2004: 103).

In the context of this volume it is also apposite to reflect on what the use of digital approaches in the health humanities means for digital humanities as a whole. The last decade has seen a good deal of reflection on the position of the digital humanities in the academy. For example, Kirschenbaum (2010) wonders about the position of the digital humanities in the study of English and emphasises its curatorial and editorial possibilities, as well as the possibility of collaboration and career development. This latter theme is also a preoccupation of Melissa Terras (2011) as she documents her own career trajectory. This search for a place in the intellectual ecosystem leads Rosenbloom to propose a model of human intellectual endeavour where digital humanities are aligned with the social sciences and are implementational and practical specialisms like the mathematical and computational foundations on which they draw. Through all of this, the drift of much thinking about the field of digital humanities is focused on archival, curatorial and editorial issues – indeed Terras (2011, 2016) conceptualises her work as a kind of heritage science and locates her work in the heritage sector.

By contrast, I would argue that the health humanities in digital guise can offer far more. Certainly, there are archival aspects, including Paul Crawford's recent AHRC project to re-examine the work of Florence Nightingale. But equally, by looking at issues such as news discourse and web-based health, we can offer a detailed archaeology of the present. Within the digital health humanities we can re-think and reconceptualise the figure of the human and address more existential questions such as 'What kind of beings do we think we are, and with what consequences?' In our own work, we have something to say about the shape and form of distress, the terms in which it is expressed under advanced liberalism. How and to what extent is this medicalised in popular discourse? How is it that sexuality and reproductive health have become sites for tutelary interventions involving advice, expertise and normalisation? Similarly, the shape and form of discourse about recovery in mental health reveals something about what is said, written and (arguably) thought about the trajectory of people through so called 'mental illness' and the role of creative activity in facilitating this. By exploring these issues on a larger canvas than is typically examined in qualitative discourse studies, we can examine the contours of what it means to be human in the present day. In this respect, our work in the digital health humanities goes beyond curatorial concerns and enables us to ask larger scale questions about the human condition itself. There is much more to the digital health humanities than a focus on curation and heritage; we need not – indeed cannot – take the figure of the human unquestioned. We can re-think and re-pattern it.

Further reading

- 1 Crawford, P. Brown, B., Baker, C., Tischler, V. and Abrams, B. (2014). *Health humanities*. London: Palgrave.
Our recent manifesto for the field of health humanities. This shows what we are attempting to do in the field, and covers topics such as the performing arts, narrative, literature, visual art and anthropology. It explains how our vision is different from that typically found in the medical humanities.
- 2 Harvey, K., Brown, B., Crawford, P., Macfarlane, A. and McPherson, A. (2007). 'Am I normal?': Teenagers, sexual health and the internet. *Social Science & Medicine* 65: 771–781.
Our original study of questions written to a website dealing with teenagers' health concerns. It shows the combination of corpus-based techniques to identify the major contours of a dataset and the drill-down into discourse analysis to illustrate the shape and form of the language.
- 3 Jones, T., Wear, D. and Friedman, L.D. (2014) The health humanities reader.
This is a book of edited chapters which surveys the rich body of work that has emerged from the field – and seeks to imagine fresh approaches to the health humanities. The collection's contributors reflect the extraordinary diversity of the field, including scholars from the disciplines of disability studies, history, literature, nursing, religion, narrative medicine, philosophy, bioethics, medicine and the social sciences.
- 4 Atanasova, D., Koteyko, N. Brown, B. and Crawford, P. (2019). Representations of mental health and arts participation in the national and local British press, 2007–2015. *Health* 23(1): 3–20.
Our study of press coverage of the arts, mental health and recovery, referred to in this chapter, with further details of the analysis and findings.

References

- Acquisti, A., Brandimarte, L. and Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science* 347(6221): 509–514.
- Adolphs, S., Brown, B., Carter, R., Crawford, P. and Sahota, O. (2004). Applying corpus linguistics in a health care context. *Journal of Applied Linguistics* 1(1): 9–28.
- Ahlzen, R. (2007). Scientific contribution: Medical humanities-arts and humanistic science. *Medicine, Health Care and Philosophy* 10: 385–393.
- Aragonès, E., López-Muntaner, J., Ceruelo, S. and Basora, J. (2014). Reinforcing stigmatization: Coverage of mental illness in Spanish newspapers. *Journal of Health Communication* 19(11): 1248–1258.
- Atanasova, D., Koteyko, N., Brown, B. and Crawford, P. (2019). Representations of mental health and arts participation in the national and local British press, 2007–2015. *Health* 23(1): 3–20.
- Baker, C., Crawford, P., Brown, B., Carter, R. and Lipsedge, M. (2010). *Madness in post 1945 British and American fiction*. London: Palgrave.
- Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.
- Baker, S., Brown, B. and Davies, H. (2013). We always invite residents to come along . . . *Discourses of Citizenship Among Local Government Stakeholders, Contemporary Wales* 26: 205–223.
- Bakhtin, M.M. (1981). *The dialogic imagination: Four essays* (M. Holquist, ed., C. Emerson and M. Holquist, trans.). Austin, TX: University of Texas Press.
- Bates, V., Bleakley, A. and Goodman, S. (eds.). (2014). *Medicine, health and the arts: Approaches to medical humanities*. Oxford: Routledge.
- Bednarek, M. (2016). Voices and values in the news: News media talk, news values and attribution. *Discourse, Context and Media* 11: 27–37.
- Betton, V. and Tomlinson, V. (2013). Social media can help in recovery – but are mental health practitioners up to speed? *Mental Health and Social Inclusion* 17(4): 215–219.
- Bilić, B. and Georgaca, E. (2007). Representations of “mental illness” in Serbian newspapers: A critical discourse analysis. *Qualitative Research in Psychology* 4(1–2): 167–186.

- Bleakley, A. and Marshall, R. (2013). Can the science of communication inform the art of the medical humanities? *Medical Education* 47(2): 126–133.
- Bolton, G. (2008). Boundaries of humanities: Writing medical humanities. *Arts and Humanities in Higher Education* 7(2): 131–114.
- Boulton, A. and Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning* 67(2): 348–393.
- Brawer, J.R. (2006). The value of a philosophical perspective in teaching the basic medical sciences. *Medical Teacher* 28(5): 472–474.
- Brown, B. (2015). Towards a critical understanding of mutuality, nursing and social capital. *Journal of Psychiatric and Mental Health Nursing* 22: 829–835.
- Brown, B. and Baker, S. (2012). *Responsible citizens: Individuals, health and policy under neoliberalism*. London: Anthem Press.
- Brown, B., Crawford, P. and Carter, R. (2006). *Evidence based health communication*. Maidenhead: Open University Press.
- Burman, J.T., Green, C.D. and Shanker, S. (2015). On the meanings of self-regulation: Digital humanities in service of conceptual clarity. *Child Development* 86(5): 1507–1521.
- Burns, J.A. (2016). Role of the information professional in the development and promotion of digital humanities content for research, teaching, and learning in the modern academic library: An Irish case study. *New Review of Academic Librarianship* 22(2–3): 238–248.
- Bury, M. (2001). Illness narratives: Fact or fiction? *Sociology of Health & Illness* 23(3): 263–285.
- Calsamiglia, H. and López Ferrero, C. (2003). Role and position of scientific voices: Reported speech in the media. *Discourse Studies* 5(2): 147–173.
- Carletti, L. (2016). Participatory heritage: Scaffolding citizen scholarship. *International Information & Library Review* 48(3): 196–203.
- Chamberlain, J.M. (2016). *Medicine, risk, discourse and power*. London: Routledge.
- Charon, R. (2017). To see the suffering. *Academic Medicine* 92(12): 1668–1670.
- Christakis, N.A. (1995). The similarity and frequency of proposals to reform US medical education: Constant concerns. *Journal of the American Medical Association* 274(9): 706–711.
- Cipriano, A., Cella, S. and Cotrufo, P. (2017). Nonsuicidal self-injury: A systematic review. *Frontiers in Psychology* 8: 1946. doi:10.3389/fpsyg.2017.01946.
- Clough, P.T. (2016). Rethinking race, calculation, quantification, and measure. *Cultural Studies – Critical Methodologies* 16(5): 435–441.
- Cochrane, D. (1989). Poverty, probation and empowerment. *Probation Journal* 36(4): 177–182.
- Conway, M. and O'Connor, D. (2016). Social media, big data, and mental health: Current advances and ethical implications. *Current Opinion in Psychology* 9: 77–82.
- Coverdale, J.H. and Nairn, R. (2006). A research agenda concerning depictions of mental illness in children's media. *Academic Psychiatry* 30(1): 83–87.
- Crawford, P. and Brown, B. (2011). Fast healthcare: Brief communication, traps and opportunities. *Patient Education & Counselling* 82: 3–10.
- Crawford, P., Brown, B., Baker, C., Tischler, V. and Abrams, B. (2014). *Health Humanities*. London: Palgrave.
- Crawford, P., Brown, B. and Nolan, P. (1998). *Communicating care: Language in health care contexts*. Cheltenham: Stanley Thornes Publishers (A division of the Kluwer Group).
- Donohoe, M. and Danielson, S. (2004). A community-based approach to the medical humanities. *Medical Education* 38(2): 204–217.
- Dysart-Gale, D. (2008). Lost in translation: Bibliotherapy and evidence-based medicine. *Journal of Medical Humanities* 29: 33–43.
- Edelstein, D. (2016). Intellectual history and digital humanities. *Modern Intellectual History* 13(1): 237–246.
- Enarson, C. and Burg, F.D. (1992). An overview of reform initiatives in medical education 1906 through 1992. *Journal of the American Medical Association* 268(9): 1141–1143.

- Everett, A. (2015). The newsworthiness of mental illness: An exploration of representations of mental health and illness in the UK national press. *Journal of Promotional Communications* 3(2): 302–310.
- Favazza, A. (1996). *Bodies under siege: Self-mutilation and body modification in culture*. Baltimore: The John Hopkins University Press.
- Fletcher, E.H. (2016). Dis/assembling schizophrenia on YouTube: Theorizing an analog body in a virtual sphere. *Journal of Medical Humanities* 37: 257–274.
- Flis, I., Steinová, E. and Wouters, P. (2016). Digital humanities are a two-way street. *Isis* 107(2): 346–348.
- Fløttum, K. (2010). EU discourse: Polyphony and unclearness. *Journal of Pragmatics* 42(4): 990–999.
- Fløttum, K. and Stenvoll, D. (2009). Blair speeches in a polyphonic perspective: NOTs and BUTs in visions on Europe. *Journal of Language and Politics* 8: 269–286.
- Garden, R. (2008). Expanding clinical empathy: An activist perspective. *Journal of General Internal Medicine* 24(1): 122–125.
- General Medical Council. (1993). *Tomorrow's doctors: Recommendations on undergraduate medical education*. London: General Medical Council.
- Gilbert, P. (2009). *The compassionate mind*. London: Constable Robinson.
- Gillis, C.M. (2008). Medicine and humanities: Voicing connections. *Journal of Medical Humanities* 29(1): 5–14.
- Gjerstad, Ø. (2007). The polyphony of politics: Finding voices in French political discourse. *Critical Approaches to Discourse Analysis across Disciplines* 1: 61–78.
- Gordon, J. (2005). Medical humanities: To cure sometimes, to relieve often, to comfort always. *Medical Journal of Australia* 182(1): 5–8.
- Graham, J., Benson, L.M., Swanson, J., Potyk, D., Daratha, K. and Roberts, K. (2016). Medical humanities coursework is associated with greater measured empathy in medical students. *American Journal of Medicine* 129(2): 1334–1337.
- Greene, J.A. and Jones, D.S. (2017). The shared goals and distinct strengths of the medical humanities: Can the sum of the parts be greater than the whole? *Academic Medicine* 92(12): 1661–1664.
- Harvey, K. and Brown, B. (2012). Health communication and psychological distress: Exploring the language of self-harm. *Canadian Modern Language Review* 68(3): 316–340.
- Harvey, K., Brown, B., Crawford, P., Macfarlane, A. and McPherson, A. (2007). 'Am I normal?' Teenagers, sexual health and the Internet. *Social Science & Medicine* 65: 771–781.
- HM Government. (2011). No health without mental health. *HM Government*. Available at: www.gov.uk/government/uploads/system/uploads/attachment_data/file/213761/dh_124058.pdf (Accessed 4 April 2016).
- Holley, L.C., Tavassoli, K.Y. and Stromwall, L.K. (2016). Mental illness discrimination in mental health treatment programs: Intersections of race, ethnicity, and sexual orientation. *Community Mental Health Journal* 52(3): 311–322.
- Horne, O. and Csipke, E. (2009). From feeling too little and too much, to feeling more and less? A nonparadoxical theory of the functions of self-harm. *Qualitative Health Research* 19: 655–667.
- Houston, B.J., Spialek, M.L. and Perreault, M.F. (2016). Coverage of posttraumatic stress disorder in the New York Times, 1950–2012. *Journal of Health Communication* 21: 240–248.
- Jain, V.K., Kumar, S. and Fernandes, S.L. (2017). Extraction of emotions from multilingual text using intelligent text processing and computational linguistics. *Journal of Computational Science* 21: 316–326.
- Kemp, S.J. and Day, G. (2014). Teaching medical humanities in the digital world: Affordances of technology-enhanced learning. *Medical Humanities* 40: 125–130.
- Kirschenbaum, M.G. (2010). What is digital humanities and what's it doing in English departments? *Association of Departments of English Bulletin* 150: 55–61.
- Kleinman, A. (1988). *The illness narratives: Suffering, healing, and the human condition*. New York: Basic Books.
- Klin, A. and Lemish, D. (2008). Mental disorders stigma in the media: Review of studies on production, content, and influences. *Journal of Health Communication* 13: 434–449.

- Koltay, T. (2016). Library and information science and the digital humanities. *Journal of Documentation* 72(4): 781–792.
- Lee, H., Warschauer, M. and Lee, J.H. (2017). The effects of concordance-based electronic glosses on L2 vocabulary learning. *Language Learning and Technology* 21(2): 32–51.
- McCarthy, M. and Handford, M. (2004). ‘Invisible to us’: A preliminary corpus-based study of spoken business English. In U. Connor and T. Upton (eds.), *Discourse in the professions: Perspectives from corpus linguistics*. Amsterdam: John Benjamins.
- McGinty, E.E., Kennedy-Hendricks, A., Choksy, S. and Barry, C.L. (2016). Trends in news media coverage of mental illness in the United States: 1995–2014. *Health Affairs* 35(6): 1121–1129.
- Meraz, S. (2011). The fight for ‘how to think’: Traditional media, social networks, and issue interpretation. *Journalism* 12(1): 107–127.
- Mirza, M. (2006). The arts as painkiller. In M. Mirza (ed.), *Culture vultures: Is UK arts policy damaging the arts?* London: Policy Exchange, pp. 93–110.
- Montgomery, M. (2007). *The discourse of broadcast news: A linguistic approach*. Abingdon: Routledge.
- Motschenbacher, H. (2016). A corpus linguistic study of the situatedness of English pop song lyrics. *Corpora* 11(1): 1–28.
- Olstead, R. (2002). Contesting the text: Canadian media depictions of the conflation of mental illness and criminality. *Sociology of Health and Illness* 24(5): 621–643.
- Olthuis, G. and Dekkers, W. (2003). Medical education, palliative care and moral attitude: Some objectives and future perspectives. *Medical Education* 37: 928–933.
- Parker, J. (2018). Stories, narratives, scenarios in medicine. *Arts and Humanities in Higher Education* 17(1): 3–19.
- Petersen, A., Bleakley, A., Brömer, R. and Marshall, R. (2008). The medical humanities today: Humane health care or tool of governance? *Journal of Medical Humanities* 29(1): 1–4.
- Plappert, G. (2017). Candidate knowledge? Exploring epistemic claims in scientific writing: A corpus-driven approach. *Corpora* 12(3): 425–457.
- Ross, M. (1994). Maggy Ross. In L. Pembroke (ed.), *Self-harm: Perspectives from personal experience*. London: Survivors Speak Out.
- Royles, D. (2016). Teaching digital humanities with oral history: The staring out to sea oral history project and OHMS in the DH classroom. *The Oral History Review* 43(2): 408–420.
- Schudson, M. (2003). *The sociology of the news*. New York: W W Norton & Company.
- Schwarz, M.R. and Wojtczak, A. (2002). Global minimum essential requirements: A road towards competence-oriented medical education. *Medical Teacher* 24: 125–129.
- Scott, M. (2016). *WordSmith tools version 7*. Stroud: Lexical Analysis Software.
- Seale, C. (2006). Gender accommodation in online cancer support groups. *Health* 10: 345–360.
- Seale, C., Boden, S., Williams, S., Lowe, P. and Steinberg, D. (2007). Media constructions of sleep and sleep disorders: A study of UK national newspapers. *Social Science & Medicine* 65: 418–430.
- Showalter, E. (1985). *The female malady: Women, madness, and English culture, 1830–1980*. New York: Pantheon Books.
- Suzuki, L. and Calzo, J. (2004). The search for peer advice in cyberspace: An examination of online teen bulletin boards about health and sexuality. *Applied Developmental Psychology* 25: 685–698.
- Svensson, P. (2010). The landscape of digital humanities. *DHQ: Digital Humanities Quarterly* 4(1). Available at: <http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html> (Accessed 22 January 2016).
- Terras, M. (2011). Present, not voting: Digital humanities in the Panopticon: Closing plenary speech, digital humanities conference 2010. *Literary and Linguistic Computing* 26(3): 257–269.
- Terras, M. (2016). A decade in digital humanities. *Journal of the Siberian Federal University: Humanities & Social Sciences* 7: 1637–1650.
- van Heek, J., Arning, K. and Zieffle, M. (2017). The surveillance society: Which factors form public acceptance of surveillance technologies. *Communications in Computer and Information Science* 738: 170–191.

- van Leeuwen, T. (1996). The representation of social actors. In C.R. Caldas-Coulthard and M. Coulthard (eds.), *Text and practice*. London: Routledge, pp. 32–70.
- Widder, J. (2004). The origins of medical evidence: Communication and experimentation. *Medicine, Health Care and Philosophy* 7: 99–104.
- Wodak, R. (2009). Critical discourse analysis: History, agenda, theory, and methodology. In R. Wodak and M. Meyer (eds.), *Methods for critical discourse analysis*, 2nd revised ed. London: Sage, pp. 1–33.
- Wooffitt, R. (1992). *Telling tales of the unexpected*. Brighton: Harvester Wheatsheaf.
- Wooliscroft, J.O. and Phillips, R. (2003). Medicine as performing art: A worthy metaphor. *Medical Education* 37: 934–939.

English language and public humanities

'An army of willing volunteers': analysing the language of online citizen engagement in the humanities

Ben Clarke, Glenn Hadikin, Mario Saraceni and John Williams

Introduction/definitions

This chapter explores the language used in interactions among members of the public engaged in discussing topics within humanities areas – such as literature and history – as a branch of what has been termed ‘citizen science’ (where ‘science’ is intended in the broad sense of ‘knowledge’). OpenScientist.org (cited in Toerpe 2013) defines citizen science (CS) as:

the systematic collection and analysis of data; development of technology; testing of natural phenomena; and the dissemination of these activities by researchers on a primarily avocational (= hobby) basis.

CS thus involves the direct engagement of large numbers of ordinary people – citizens, volunteers – in the active co-production of knowledge. This makes it significantly different from more conventional situations where knowledge is exclusively advanced by small teams of scientists who then impart it to everybody else. As they are typically located in very different parts of the world, the participants in such interactions use English as a *lingua franca* and communicate on online forums. The discussion forums we are interested in occur on Zooniverse, a large online platform that hosts CS research projects, each of which includes its own discussion forum. As such, Zooniverse is an opportune site to inquire what role language plays in attempts to democratise knowledge.

With a view to gaining a better understanding of the discourse that unfolds in these interactions, the present authors have applied the methods of *corpus linguistics* (the analysis of language by means of large databases of authentic language) in order to address a number of related questions:

- How do newcomers negotiate their entry to and membership of such international communities? Are there any differences where non-native speakers are concerned?

- What are the linguistic implications when the methods of CS are transposed to a humanities setting?
- What does the language of the forums reveal about the ways in which volunteers position themselves in relation to professional experts?
- By what methods can such ‘revealing’ language be identified? At what linguistic level should the search be conducted – individual words, phrases or longer textual segments?
- Does the use of language by citizen scholars change over time as and when they become acculturated into their online community?

In responding to these questions, the authors deploy a number of theoretical concepts which are fully defined in the text as they arise. These include *community of practice*, *lexical selection* and *semantic sequence*. The methods of corpus linguistics are explained in more detail in the section concerning main research methods. The ‘Current contributions and research’ section gives a flavour of the different, but nevertheless ultimately converging, directions in which our research has taken us, while the section ‘Sample analysis’ offers a fine-grained corpus analysis of the words *I* and *we* as they occur in one of the Zooniverse forums. Finally, some further possibilities for research are explored based on the Zooniverse data, as well as the crowdsourcing methodology it embodies.

Background

Changing forms of participation: from top-down to citizen science

One of the most fundamental sociocultural changes that have taken place in the first two decades of this century relates to the production and diffusion of knowledge. In the past, these were the remit of very restricted groups of people. Scientists and scholars conducted research and published their findings in specialised journals, generally available only to their academic peers through their university libraries. Then there was another small circle of people – journalists, popular science writers, TV documentary presenters, etc – who popularised some of those findings for the general public. The general public, in turn, were very much in a position of consumers of knowledge, in a system which was strictly unidirectional (Figure 28.1).

Not only were the general public in an entirely passive position in the knowledge flow, but the content that they were able to access was also relatively static, as it was available only through print and traditional broadcast media.

With the growing diffusion of the Internet in the late 1990s and, especially, with the wide availability of fast, always-on broadband technology in the new century, things began to change rapidly. Unlike traditional media, the new Internet-based ones were

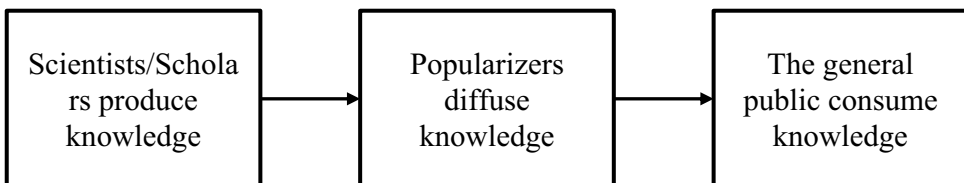


Figure 28.1 The traditional model for the production and diffusion of knowledge

increasingly more interactive. The label Web 2.0, which became popular in the early 2000s, refers precisely to the transition between a more passive mode of utilisation (e.g. reading a newspaper, on- or offline) to one where people were able to interact not only with one another but also with those who produce content. This, in turn, triggered further development in the same direction, culminating more recently in the proliferation of so-called social media, such as Facebook, Twitter, YouTube and so on – large online platforms where content is not just consumed but also actively produced by their participants.

The content thus simultaneously consumed and produced is not limited to social purposes (e.g. keeping in touch with friends), but also involves – and has radically changed – the knowledge flow outlined earlier. In particular, two important variables are now different. First, science, intended in the broadest and original sense of the word as knowledge, is no longer exclusively produced by a restricted circle of scientists: ordinary people are increasingly able to participate in the process of producing it. Second, the flow is no longer unidirectional but bi- and multi-directional. A good example to illustrate this change is the shift from traditional print manuals to online support forums for the use and maintenance of products such as cameras, mobile phones, software and so on. Whereas previously people would typically rely entirely on the user manual that would come with a product, it is now increasingly common for users to participate in online forums in order to access community-based knowledge in the form of advice, tips, hands-on experience, tutorials, etc., so that everybody can not only benefit from but also contribute to producing. Knowledge is therefore both accessed more easily and co-produced.

The co-production of knowledge is the core principle of the activity that has come to be called ‘citizen science’. The word ‘citizen’ signifies both the ‘ordinary person’ roles of its participants and the democratic nature of the system. Possibly the most famous and most widely used CS platform is the online encyclopaedia Wikipedia, which is based on the principle that all its content is produced, freely accessible, editable and shared by its own users. Within this system, nobody is inherently more knowledgeable than others, and the roles defined in Figure 28.1 are profoundly transformed (Figure 28.2).

Within this new system, each citizen is both a producer and a consumer of knowledge, while the middleman role of the populariser is superfluous. Importantly, citizens are engaged in the production of knowledge not in isolation but in communication with one another, in a concerted effort to ensure that content is continuously as up to date as possible.

Zooniverse

One of the largest online CS platforms based on this principle is Zooniverse (www.zooniverse.org). Zooniverse is a collection of large-scale research projects involving ordinary people, largely unpaid volunteers, engaged in data analysis aimed at what might loosely be termed ‘pattern recognition’. Typically, a Zooniverse participant analyses a number of images and identifies aspects of their content. Depending on the specific focus of each project, this may range from galaxies to animals, old manuscripts, minerals and many more. Hundreds of people might look at the same object, its ultimate identification depending on the weight of opinion arising from these observations. Benefitting thus from the ‘wisdom of the crowd’, Zooniverse projects aim to analyse information ‘more quickly and accurately than would otherwise be possible’ (Zooniverse.org 2017). Thus, at this level, CS may be characterised as the ‘crowdsourcing’ of data analysis.

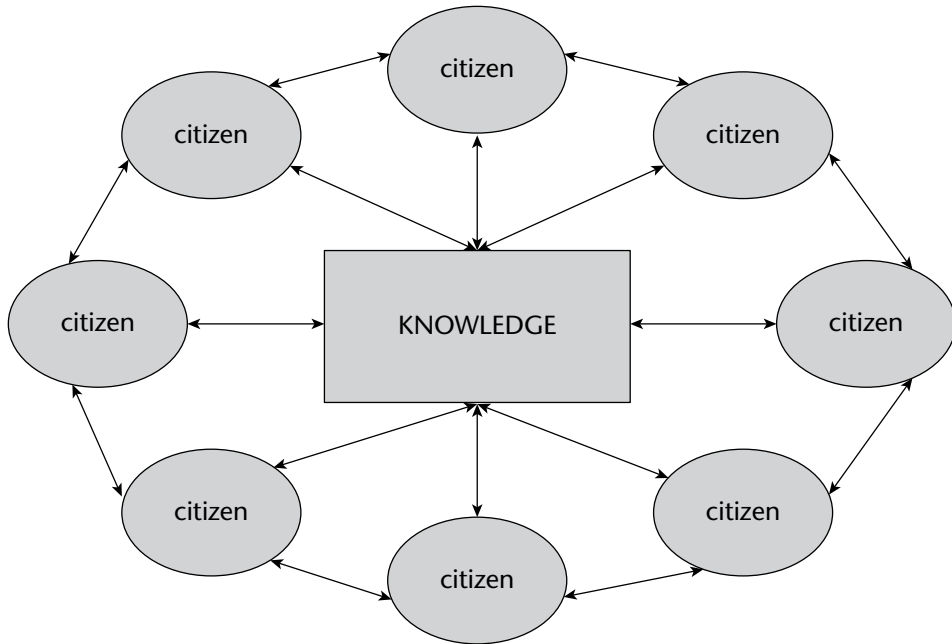


Figure 28.2 The 'citizen' model for the production and diffusion of knowledge

A key component of each Zooniverse project is the discussion board,

where volunteers can work together with each other and with the research teams. These boards are integrated with each project to allow for everything from quick hashtagging to in-depth collaborative analysis. . . . Many of the most interesting discoveries from Zooniverse projects have come from discussion between volunteers and researchers.

(Zooniverse.org 2017)

It is the analysis of the language used on these discussion boards that forms the basis of the current authors' research. It is envisaged that such an analysis can shed light on the motivations, preoccupations and trajectories of citizen scientists, as well as the discursive processes that contribute to the co-construction of knowledge in a CS setting.

The Zooniverse humanities projects: citizen science as a model?

At the time of this writing, there were five Zooniverse projects that could properly be described as 'humanities' projects, relating to records of human activity and human history, rather than, say, astronomical phenomena or animal behaviour. The five projects are listed here, together with a brief description of each:

- Ancient Lives – transcribing ancient Greek papyri [Currently inactive]
- Anno Tate – transcribing artists' diaries, letters, and sketchbooks
- Measuring the Anzacs – transcribing New Zealand soldiers' accounts of World War I
- Operation War Diary – transcribing British soldiers' diaries from World War I
- Shakespeare's World – transcribing documents and recipes written by Shakespeare's contemporaries

As can be seen, all five projects are concerned with the transcription and deciphering of ‘difficult-to-read’ documents and manuscripts. In this respect, they are no different from some other Zooniverse projects – such as Old Weather and Notes from Nature – which use old documents as sources of historical data about the physical and natural world, with a view to enhancing understanding of current phenomena such as climate change and its effects on habitats.

From this perspective, there seems to be a methodological convergence between CS and forms of citizen engagement in humanities research. This convergence is centred on the idea of research as an empirical activity of examining and classifying data. At one level, this approach is very much in tune with the mission of the digital humanities, with its focus on ‘preserving physical artifacts’, systematic record-keeping and quantification, for example, in regard to archaeological data (Schreibman et al. 2004). Such painstaking preoccupation with data in humanities is not without its critics, such as Allington et al. (2016), who see it as side-lining the essential activities of interpretation and criticism that give shape – and a potentially radical edge – to the human sciences. Some radical ‘digital humanists’, while acknowledging the contribution of the natural sciences, quantitative approaches and the ‘custodial’ responsibilities of humanities scholars, are calling for a ‘second wave’ of digital humanities which would be ‘qualitative, interpretive, experiential, emotive, generative in character’ (Humanities Blast 2011). In so doing, they are absorbing and re-appropriating the sorts of criticisms articulated by Allington et al. (2016). It is unclear to what extent such aspirations are filtering through to the ‘citizen humanists’ who are active in the Zooniverse forums.

Not that citizen scientists in the physical and natural sciences are necessarily immune to concerns about their role and practices as researchers. In effect CS (and its humanities analogues) depend, for statistical reliability, on the observations of large numbers of unpaid enthusiasts, frequently described as an ‘army of willing volunteers’ (Brain Bank North West 2015; Winder 2012) ‘doing the grunt work of reviewing data for free’ (Martoccia 2012). This metaphor occasionally finds an echo in the forum contributions for the military-related Zooniverse projects:

Perhaps this is intended to familiarise us with the experiences of the ‘tommies’, i.e. kept in the dark and expected only to do as much work as possible, without question or reward. (Operation War Diary)

Indeed, some critics have expressed disquiet at the idea of CS as ‘free labour’ in the interests of the greater glory of a relatively small number of professional scientists (Illingworth and Allen 2016: 5–12). It remains to be seen how such tensions are played out linguistically in the forum data for the Zooniverse humanities projects.

Critical issues and topics

Recruitment and retention of crowdsourcing volunteers

The original impetus for taking an interest in the language used by citizen scientists (and humanists) in online forums came from the so-called Volunteer and Crowdsourcing Economics Research (VOLCROWE) project (www.volvrowe.org), a team of researchers interested in fostering ‘sustainable volunteer engagement’ in ‘digital volunteering’ through an investigation into the motivations of Zooniverse contributors and the factors which lead to

the success or otherwise of individual Zooniverse projects. The success of a project depends on the project managers' ability:

- To recruit and retain large numbers of volunteers
- To communicate with them effectively so that they can fulfil their role
- To encourage interaction and discussion so that they are engaged and committed

It is easy to see why these issues are of interest to linguists. If the volunteers are to be recognised as partners alongside their professional peers, then they need to be able to 'speak the same language'. One scenario postulates a levelling-up of the language used in the online discussions, whereby new 'citizen' participants gradually acquire the relevant technical vocabulary and discourse conventions deployed by the professional researchers. However, there is a potential double-bind: as the language levels up, this may in turn act as a deterrent to later newcomers. Over time, the overall level of participation may plateau or even fall, gradually becoming confined to a small circle of 'initiates'. Thus, an alternative scenario is envisaged, whereby the professionals adjust their language to the level of knowledge of the lay participants. This would favour an overall increase in the level of participation but at the risk of reinforcing boundaries between professionals and 'citizens'.

Communities of practice

In order to conceptualise the relationships and patterns of interaction between the various participants in the Zooniverse discussion boards, the present authors draw on the notion of a 'community of practice'. This term refers to 'a group of people who interact, learn together and build relationships, and in the process develop a sense of belonging and mutual commitment' (Wenger et al. 2002: 34). Though cautious about applying the term to any and every virtual 'meeting place' on the Internet, Mercer (2000: 119–120) does concede that some online communities do satisfy the criteria of 'specialised knowledge' and 'common goals' that define a community of practice, citing the example of a rock band fan club. We would suggest that the Zooniverse communities qualify on the same criteria.

Communities of practice are characterised by three fundamental traits:

- 1 **The domain:** A community of practice . . . has an identity defined by a shared domain of interest. Membership therefore implies a commitment to the domain, and therefore a shared competence that distinguishes members from other people.
- 2 **The community:** In pursuing their interest in their domain, members engage in joint activities and discussions, help each other, and share information. They build relationships that enable them to learn from each other; they care about their standing with each other.
- 3 **The practice:** Members of a community of practice are practitioners. They develop a shared repertoire of resources: experiences, stories, tools, ways of addressing recurring problems – in short a shared practice. This takes time and sustained interaction.

(Wenger-Trayner and Wenger-Trayner 2015: 2)

A community of practice typically includes 'older' members with greater experience of the social practices established within the community, and newer ones, who might still be negotiating those conventions. This also means that the more experienced members may often act as mentors for the novices. In this sense, regular interaction and communication

plays a key role, since ‘[t]he social relationship between . . . established practitioners and novices . . . facilitates the transfer of knowledge and ensures continuity within the community’ (Manzoor 2016: 266).

This brief description of communities of practice does appear to fit the constitution and the social dynamics of Zooniverse discussion forums.

Roles within Zooniverse communities

The main Zooniverse website states that:

You don’t need any specialised background, training, or expertise to participate in any Zooniverse projects. We make it easy for anyone to contribute to real academic research, on their own computer, at their own convenience.

However, the site also makes clear that ‘volunteers – hundreds of thousands of people around the world . . . come together to assist professional researchers’. Thus, despite the *citizen science* principle guiding the platform, roles are still fairly clearly defined, distinguishing professional researchers from ordinary people volunteering to contribute their enthusiasm and time. This means that although specialised training is not required, there is inevitably an implicit assumption that participants in a particular project will have *some* expertise in the subject. Therefore, new members are aware that they are joining a community where (1) there are professional researchers who lead the project and (2) existing members are likely to be, to some degree, experts both in their chosen topic and in the communicative/social practices established within the community.

Within this frame, one of the areas that it is interesting to investigate is how the roles of mentor/expert members and mentee/novice are linguistically negotiated. This issue will be explored more fully later, with specific reference to the international dimension of Zooniverse communities.

The diachronic dimension

Since, as with most computer-mediated communication (CMC), Zooniverse forum posts are labelled with usernames, dates and times, it is possible to follow the development of community-specific language as it moves through the forums and it is transmitted from user to user. It is proposed that such language plays an important part in building the in-group identity essential to communities of practice.

Furthermore, individual linguistic trajectories within Zooniverse become open to investigation. Some new contributors may decide CS is not for them and drop out very quickly, or be active only for a short time on a single project or thread. Others, however, will become long-term active members of the community and potentially across many different projects. As they acquire the techniques of data analysis and earn ‘*kudos*’ within the community, their language may change. These relatively short-term diachronic changes are amenable to analysis along many parameters and according to many phenomena.

As can be inferred from other sections of this chapter, one of the primary concerns of our team’s research has been language that is indicative of matters of an interpersonal kind, where intersubjectivity broadly conceived (e.g. to negotiate; to persuade; to enlighten; to share experiences, opinions, evaluations, etc; to position others with respect to ourselves) shapes and is shaped by certain encodings in language (e.g. grammatical-mood, modality,

speech acts). Thus, in the Zooniverse forums, two primary patterns of user behaviour across time have so far been observed: (1) those users who in their earlier time periods on the forum use a greater proportional frequency of weak forms of modal commitment (e.g. *possibly*, *may*, *I think*), which decrease as time goes on, with these users also often initially deploying proportionally few strong forms of modal commitment (e.g. *must*, *certainly*), which therefore proportionally increase as they spend more time on the forum, and (2) those users who adopt weaker forms of modal commitment when they perceive their contribution to be part of a dialogic communication and stronger modal forms when they believe their post is of a monologic type.

Lexical selection

Within an online community of practice, knowledge is transmitted, ideas circulate, roles are negotiated and community bonds are forged. This can only be done through language. Thus, questions arise as to what kinds of linguistic units are operative or relevant to these processes. These could be individual words or any of various higher-level units. Hadikin (2015) has studied the ways in which particular sequences of words establish themselves in and move through a community (whether online or offline), in a manner akin to evolutionary processes of selection. In developing his theory of ‘lexical selection’, he draws on Sinclair’s (1996) notion of the *extended lexical unit* in an attempt to identify the gene-like replicating unit that propagates itself through a community. Hadikin (2015) is careful not to give a categorical definition of an *extended selection unit* (ESU), but he cites the example of the sequence *that kind of thing* as being peculiarly characteristic of the English of the Korean community in Liverpool. Although *that kind of thing* happens to be a well-formed noun phrase, it is important to note that the replicating units can be fragmentary or cut across syntactic boundaries. Hadikin (2017) focuses on the ‘Dark Matter’ thread in the Galaxy Zoo forum on the Zooniverse site. This chapter focuses on the sequences that cluster around the fragment *we are*, such as *that we are*, *which we are*, *we are told* and *we are in*. The following analysis extends this focus to one of the Zooniverse humanities forums, while also considering the lexical environment of the personal pronoun *I*.

Computer-mediated communication as a genre (or genres)

Not surprisingly, in the wake of the Internet revolution of the 1990s and the subsequent proliferation of social media since the turn of the century, linguists have been taking an ever-increasing interest in the language of CMC in all its formats, forms and platforms. A cursory library search throws up the following book titles, among others, *Language and the Internet* (Crystal 2006), *Pragmatics of Computer-Mediated Communication* (Virtanen et al. 2013), *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives* (Herring 1996), *Virtual English: Internet Use, Language, and Global Subjects* (Enteen 2010). Within that, there is an interest in the different modes or genres of CMC, in identifying the specific linguistic features associated with particular ways of interacting online. Some noteworthy examples are Virtanen (2017) (consumer reviews), Schildhauer (2017) (personal blogs) and Zhang and Vasquez (2014) (TripAdvisor).

This last article in particular, drawing on the work of the genre analyst John Swales, recalls the ‘classic’ analyses of more formalised written genres, especially academic writing, associated, alongside Swales, with such names as Ken Hyland and Chris Tribble. (See for example Swales 1990.) Some of this work focuses on the identification of so-called

‘semantic sequences’, thought to be characteristic of particular formal or academic genres. These ‘semantic sequences’ are in some ways reminiscent of Hadikin’s (2015) lexical ‘memes’ (see earlier), but tend to be longer and allow for more variation in the particular lexical items that can fill the slots in the sequence. According to Hunston (2008: 271),

[s]emantic sequences are recurring sequences of words and phrases that may be very diverse in form and which are therefore more usefully characterised as sequences of meaning elements rather than as formal sequences.

One example of a semantic sequence from scientific writing, identified by Gledhill (2000: 152) is:

the fact that + **NEGATIVE EMPIRICAL OBSERVATION** + EXPLANATION

The fact that | **this enhancement does not occur in females** | implies that such onco-genes were not involved

The fact that | **sequential accumulation of LOH was not observed** | might be due to early monitoring

Groom (2017), mindful of differences in disciplinary writing conventions between science and humanities, has proposed a number of semantic sequences associated with academic writing in the humanities, for example:

CONCEPTUALISER + CONCEPTUALISING PROCESS + **DISCIPLINARY ENTITY** + *as* + CONCEPTUALISATION

we | can see | **the moment of revelation** | *as* | a moment of alienation and misery

Mark Storey | describes | **the scene** | *as* | mundane

With the Zooniverse website bringing together both science and humanities projects under one roof as it were, one can see how a possible research agenda might take shape, essentially comparative in nature, centred around the identification of lexical units or semantic sequences identified as being characteristic of online ‘citizen’ discussion, either within or across academic disciplines. What work are they doing in terms of knowledge production or community cohesion? It should be noted in this context, however, that there is already some existing work in digital humanities – albeit with a rather different purpose – that is concerned with identifying recurring phrases and ‘character strings’ in a given discourse type – see Winters’ (2017) account of her study of UK parliamentary records (Hansard).

Main research methods: corpus linguistics

Linguistics is regarded as a strategically important discipline within the humanities, because of the capacity of language to be both ‘informational’ and ‘social’ (Rosenbloom 2012). Computers are very good at counting and manipulating items of data. Thus, it is not surprising that language corpora, which treat language as data, have become a widely acknowledged component of the digital humanities toolkit (Wooldridge 2004; Allington et al. 2016; Humanities Blast 2011; Terras 2010), as they enable reliable evidence-based judgements to be made about linguistic phenomena. For those unfamiliar with corpora, John Sinclair, who may be considered the ‘father’ of corpus linguistics, offers the following definition:

A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.

(Sinclair 2005)

In the present case, the ‘language variety’ is the language of the contributors to the Zooniverse forums and the phrase ‘external criteria’ refers to the fact that the selection of data to be included in the corpus is not made on the basis of the linguistic phenomena the researchers wish to observe. In order for conclusions to be valid, the data from the forums need to be as complete as possible, limited only by time, practical and technical constraints.

The Zooniverse forum threads are available on the Internet for all to read, but – due to their dynamic nature and the variety of forum designs – it proved something of a technical challenge to download the entire set in a corpus-friendly format, tagged with usernames and timestamp data. Nevertheless, the technical challenge was overcome, and the resulting corpus runs to around 14 million words covering 43 different projects. The five humanities projects taken together represent just over 4.5 percent of the total word count. The forum data from each project are stored as a separate sub-corpus, but it is possible to combine sub-corpora for analysis and also to make sub-corpora of individual users’ contributions across different projects.

The corpus or sub-corpora can be interrogated using a variety of different software packages. (Hence the rather approximate figures in the previous paragraph – different software counts words in different ways.) Most of the analyses in this chapter were carried out using the Sketch Engine website (Sketch Engine n.d.), the free, simple-to-use software AntConc (Anthony 2016), or – particularly the fully worked analysis in the section later – the powerful WordSmith Tools package (Scott 2017). All three packages allow the user to generate a variety of different outputs from the data including:

- *Frequency listing* – a list of all the words in the corpus sorted according to frequency; this will vary from corpus to corpus but, for any kind of natural language, the top positions in the list will be dominated by the common closed-class grammar and function words such as *the*, *of*, *and*, and *not*
- *Concordance* – a list of examples of the search term in context (see Figures 28.4 and 28.6); concordances can be sorted according to the words occurring to the left or the right of the search term, which enables collocational patterns to be shown
- *Collocations listing* – a list of the words most frequently found in proximity to the search term; a variety of statistical and display options are available according to the software
- *Keywords listing* – this is discussed later

Corpus-based versus corpus-driven

There are different principled methodological approaches to exploiting corpus data. The first, corresponding to a classical position in the philosophy of science, involves deriving general statements and hypotheses – initially from outside the corpus – which are then iteratively tested against data. This is known as a *corpus-based* approach. Aarts’s (1991) research into English grammar is a good example:

We start from an intuition-based grammar consisting of a set of well-established rules, and confront this grammar with the material contained in a corpus. This confrontation

will lead to an expansion and revision of the original, intuition-based grammar, which in the course of this process becomes an observation-based grammar.

(Aarts 1991: 58)

The *corpus-driven* approach, on the other hand, rejects – as far as is humanly possible – any form of intuition and treats the data as a starting point, using the available software to tease out patterns that may then become the basis for further investigation.

Credit for naming these two approaches is usually attributed to Tognini-Bonelli (2001) but in fact the distinction, in all but name, is an old one in the philosophy of science, corresponding to the debate between Carnap and Popper:

Carnap's verification is from the bottom up: make observations and see how they add up to confirm or verify a more general statement. Popper's falsification is from the top down. First form a theoretical conjecture, and then deduce consequences and test to see if they are true.

(Hacking 1983: 3)

In terms of corpus linguistics, Carnap's position would be regarded as corpus-driven and Popper's as corpus-based.

In practice, both approaches co-exist, usually peacefully nowadays, and the choice of one or the other might be quite opportunistic depending on the interests of the researcher or the particular phase of their research. The corpus-based approach can get to the heart of an issue relatively quickly, whereas being corpus-driven can be useful when the researcher is faced with an unfamiliar dataset and does not quite know what questions they should be asking.

In terms of the analyses presented in the following sections, 'Sorry for my English' may be considered corpus-based, as it starts from a predetermined set of search items. By contrast, the current research discussed is much more corpus-driven insofar as the initial objects of attention (the words *or* and *I*) are thrown up by the software as being particularly significant, without any 'cherry-picking' on the part of the researchers. The next section presents an important technique that is characteristic of corpus-driven research.

Keyword analysis

In AntConc, as well as in other corpus analysis packages, keywords are defined as words that are statistically more frequent in one corpus than in another (Anthony 2006). In order to generate a list of keywords, it is necessary to define two corpora (or sets of corpora) as a research corpus and a reference corpus respectively. The keyword listing will display those words that are statistically 'key' in the research corpus as compared to the reference corpus. The keywords can be ordered according to two different measures of 'keyness', namely log-likelihood or chi-square value.

In addition to keywords, the WordSmith Tools package is able to compute 'key clusters', that is, statistically significant sequences of two or more words. By looking at key clusters alongside keywords, the corpus linguist is more able to make informed decisions about what is the most promising unit of analysis – individual words or longer recurring sequences (see also the section concerning lexical selection earlier).

The systematic analysis of keywords (and key clusters) is a feature of much corpus-driven research (see the section earlier). They are not chosen *a priori* by the researcher but can be very useful for identifying the characteristic linguistic patterns of a given

corpus. One initial method of proceeding is to classify the keywords according to Scott's (1999) threefold categorisation:

- 1 **Proper nouns** – Different people, places, etc., will be named in different corpora, even corpora that are comparable in other regards.
- 2 **'Aboutness' (or 'Domain') words** – These are open-class lexical words specific to the domain represented by the research corpus, in cases where the reference corpus represents a different domain or is a general corpus covering a multitude of domains. For example, in the Zooniverse humanities corpus, *page*, *diary* and *subject* are 'Aboutness' keywords, since they refer to the business of transcribing and annotating the documents under consideration.
- 3 **'Style' words** – These are commonly used closed-class grammatical or function words, or other frequently occurring items of general vocabulary that bear no obvious relation to the corpus domain. They may appear insignificant in themselves, but if such frequently occurring items rank highly in a keyword list, they may reveal something significant about the style, register or rhetorical stance of the material in the research corpus. In the Zooniverse humanities corpus, *I*, *it* and *thanks* are examples of 'Style' keywords.

It was by means of a keyword analysis that the 'style' word *or* was identified as being particularly significant in the 'Dark Matter' corpus (see later). Similarly, both Gledhill (2000) and Groom (2017) work with style keywords (e.g. *as*) in order to tease out the characteristic semantic sequences in the discourses they study. Mahlberg (2007), in her corpus-based study of the works of Charles Dickens, focuses on five-word keyword clusters as 'pointers to local textual functions'.

Current contributions and research

English as a lingua franca on Zooniverse

As long as they have a computer and access to the Internet, Zooniverse participants – normally referred to as *zooites* – can be physically located anywhere in the world, and a typical Zooniverse community will therefore feature a wide range of language backgrounds. For this reason, on the online forums associated with each project English is used as a common language – a *lingua franca* – and the use of other languages is actively discouraged. This means that a significant number of *zooites* participate in the various discussions in a language that they do not consider their own or their main one. In this sense, the linguistic landscape of Zooniverse communities reflects that of the world at large, as English is by far the most widely used international *lingua franca*.

Roles within English as a lingua franca

The way participants position themselves with regard to the English language appears to be one of the structuring elements of the Zooniverse community of practice. Continuing the parallel between Zooniverse communities and the outside world, one can distinguish roles within the broad community of English language users in the world. Two classic categorisations are (1) the distinction between native speakers and non-native speakers of English and (2) Braj Kachru's 'three circles' of English (Kachru 1985). The concepts of so-called

‘native’ and ‘non-native’ speakers, are based on a simple demarcation line between people who grow up using English at home and in an environment where it is the main language in society and people who acquire it later in life, through education, and only use it in a restricted range of domains. Kachru’s ‘three-circles’ model takes that same principle but refines it by adding an historical and geographical element to it. This results in a three-way categorisation into:

- The ‘inner circle’, that is, the ‘cultural base’ of the language or, more simply, the countries where English is the dominant language in all aspects of society (sometimes referred to as ‘English-speaking countries’), such as the UK, Ireland, the United States, Australia and New Zealand.
- The ‘outer circle’, that is, former colonies of the British Empire and where English has remained as a colonial legacy and plays intra-national roles, for example, in administration, education, etc.
- The ‘expanding circle’, that is, countries where English has no historical presence and is only used for international communication.

While they have been repeatedly criticised in sociolinguistics (see Rampton 1990; Davies 2003; Holliday 2006, among others, for a critique of the notion of (non-)native speaker, and Bruthiaux 2003; Yano 2009; Saraceni 2010 for a critique of the ‘circles’ model), these categorisations are still rooted deeply in people’s mind-sets and influence the perception – including self-perception – of speakers of English whereby so-called ‘native speakers’, or people in the ‘inner circle’, are considered linguistically superior to so-called ‘non-native speakers’ in the other two circles.

‘Sorry for my English’

This perception manifests itself in Zooniverse forums, too, particularly through a frequently recurring piece of language behaviour whereby zooites who do not consider English their main language seem to apologise for their use of this language, primarily through the phrase ‘sorry for my English’.

Such phrases are typically found either at the very beginning of a post, immediately after the salutation or, more commonly, right at the end, and generally disconnected from the topic under discussion. For example, this is a post from the Ancient Lives forum:

can I see my transcription on a white paper, because I understand Greek and it would help to split words to solve many transcriptions by the sense.

Second question:

will I ever know if my work has been useful?

Sorry for my English, but in fact I only studied ancient Greek and Latin.

Here the participant apologises for their English and justifies the apology by pointing out that they studied other languages. The interesting question here is: in what way(s) does somebody’s English constitute an offence for which an apology is necessary?

One answer can be found in the widespread belief, in part fuelled by some traditional second language acquisition research (e.g. Long 1983; Varonis and Gass 1985), that speakers of

English as other than their first language are somehow deficient users of English and therefore bound to make mistakes. This is part of what Firth (2009) calls ‘native-speaker centric, monolingual, modernist and structuralist biases’ which produce the view that such speakers are:

perform deficient communicators who are perpetually, agonizingly, chronically struggling, like Sisyphus and his stone, to ascend the steep incline of their ‘interlanguage’, the goal being the promised land of ‘target competence’, that hallowed place reserved for the fabled and idealized native speakers.

(2009: 151)

Partly as a result of the influence of prescriptivist views about language, this ‘native-speaker-centric’ belief exists outside academia and pervades the self-perceptions of speakers of English as a *lingua franca* who feel self-conscious about their English and feel compelled to apologise about it (see also Vettorel 2014).

However, alongside its more apologetic nature, ‘sorry for my English’ can also be seen as a strategy for would-be zooites to negotiate their membership in the community. In other words, the self-deprecation of their linguistic skills can be interpreted as acting as signalling their novice status, in a way that is functionally analogous to new members’ clarifying that they are not experts and/or professionals in the subject matter of the particular project. Indeed, the linguistic apologies are often interwoven with similar comments about subject expertise, or lack thereof.

Significantly, many of the direct responses to the linguistic apologies express a very welcoming and reassuring attitude by existing forum users. This normally comes in the form of praise for the new member’s English, often accompanied by self-deprecation of one’s own poor or non-existent foreign language skills. The short exchange shown here (part of a larger one), from the Old Weather forum, gives a flavour. The user s-a-be has already signalled that (s)he is a new member:

s-a-be:

Hi,

Thanks so much for all the answers. And for the English lesson. It’s not my first language . . . And I know that’s a crappy excuse but it’s the only one I have.

Randi:

Welcome to the forum family!

Several of our top transcribers do not have English as a first language. Patience is far more important! Also, with so many sailing and navy terms, even those of us who are native speakers have problems.

Pommy Stuart:

Welcome s-a-be this is your Captain pommystuart.

I am sure your English is a lot better than my ability with your language.

Thus, it may be necessary to qualify the perception that the ‘native speaker’/‘non-native speaker’ distinction is a structuring element of the Zooniverse community of practice, by

saying that this is true only insofar as it maps onto the expert/novice distinction. Nevertheless, ‘sorry for my English’ and its close variants can be considered as one of the recurring lexical units that help to sustain community bonds.

Lexical features as evidence of wider concerns

‘Dark Matter’

The ‘Dark Matter’ corpus, consisting of a single, but lengthy thread on the Galaxy Zoo forums, was the first working corpus to be drawn from the Zooniverse data. One rationale for this choice was that the topic of the thread is discussed with equal interest both in specialist academic journals and in the non-specialist press. This offers opportunities for comparison. It was a relatively simple matter to create additional ‘Dark Matter’ corpora for the other two genres, to use as reference corpora in a keyword analysis (see earlier) of the ‘Dark Matter’ forum data. Using the keywords as pointers, the overall aim was to tease out the specific linguistic features of CMC discussion of dark matter (which distinguish it from academic and journalistic discussion) – then, further down the line, test to see if these same features were present in the other Zooniverse forums.

Once features specific to general online science discussion had been filtered out using a third reference corpus (see Williams, in preparation, for methodology), a keyword list for the forum corpus was generated. The second highest-placed ‘style’ keyword – and thus deserving of further analysis – was the simple conjunction *or*. Four discursive tendencies were identified converging around this keyword, of which three are relevant to the present context. These tendencies may be seen as indicative of the specific role of citizen scientists, their self-conception and their relationship to professional scientists. The three relevant tendencies are as follows:

1 A concern for exactitude in the use of terminology, typified by the use of the salient bigram (= 2-word string) *or rather*:

Energies or rather energy differences go proportionally as mass times electric charge to the fourth power.

This tendency extends to the self-presentation of the forum contributors:

[A]s a non physicist or rather as a post doc non physicist I have learned a useful amount of physics in the last few months.

Although this discursive move may be labelled as ‘reformulation’, it is not reformulation in the sense that complex ideas are simplified and repackaged for a lay audience (cf. Luzón 2013) – rather the opposite, reformulating ‘up’ as it were. In terms of Hunston’s semantic sequences (see earlier), the template would be something like:

LESS EXACT/TECHNICAL TERM + *or rather* + MORE EXACT/TECHNICAL TERM

The overall purpose appears to be to emphasise the contributors’ credentials as scientists, while remaining within the popularising mission of CS.

2 **Asking searching questions, exploring alternatives**

What's in the halo?; is it still boring old wimps and non baryons or is it the alloforms and/or mirror forms of hydrogen?

3 **Wilder speculation on questions fundamental to the topic of the forum, typified by the bigram *or even*:**

... different universes might co-exist or even interbreed.

These tendencies were seen as signalling an aspiration on the part of the forum contributors to go beyond the role assigned to them as mere 'data recorders' (or 'willing volunteers'), to engage in higher-order activities of interpretation and theory-building.

'Or' in the humanities forums

The next step is to test the findings about *or* from the 'Dark Matter' corpus to see if they hold true in the humanities forums. However, preliminary analysis suggests that this is not the case. The frequency of the word *or* is not so statistically significant as it was in the 'Dark Matter' forum, and the bigram *or rather* is not particularly salient at all. Most instances of *or* are narrowly connected with the task in hand and tend to fall into the following categories:

- **Uncertainty over specific transcriptions and their meanings:**

For the weights, it looks like qr for quarter and a barred L (or B?) presumably pound. It's not 100 percent clear I agree (nothing wrong with your English!), but I think where they say 'shelling more or less normal' and 'machine gun fire active during early part of evening' they are talking about the enemy firing at them, or at some part of their line.

- **Procedural queries or instructions:**

Ranks. How do I enter these. Or do I ignore them.
We're not marking the fields or transcribing these now.

- **Technical and computing issues:**

It might be to do with the way your cache clears or your browser's security settings.
[A]re you able to post comments from the Talk page? Or are comments not working at all?

There is little evidence of wilder speculation, or of a concern to display specialist knowledge or credentials. In fact, – in an echo of the 'modesty' discussed earlier – contributors sometimes use *or* with a negative to draw attention to their *lack* of expertise in the field:

I'm a biologist with zero experience reading anything like this, plus no experience with linguistics or history either, but I'm finding this project very interesting so far.
Obviously I am not a linguist or philographer.
Caveat: I don't work for IWM or the National Archives, so don't take those statements as gospel truth. :)

In schematic terms:

***I* + SELF-REFERRING VERB/PHRASE + NEGATIVE PARTICLE + PHRASE
DENOTING SPECIALISM/EXPERTISE + *or* + SECOND PHRASE DENOTING
SPECIALISM/EXPERTISE**

There are a number of possible explanations for the striking difference between the uses of *or* in the two forums. One may simply be to do with the differing subject matter; ‘dark matter’, almost by definition, cannot be observed and thus by compensation attracts a lot of speculation as to its nature and properties. Another explanation may relate to the fact that, by and large, the humanities projects are less mature in years than the big astronomical projects like Galaxy Zoo. There may be a higher proportion of inexperienced contributors, still ‘feeling their way’ in terms of technical procedures. Alternatively, it may be the case that volunteers with an interest in the humanities tend to be less at ease than their scientific peers with highly disciplined routines for recording observations and analysing data. Further studies will be necessary before these questions can be definitively answered.

Sample analysis: the case of ‘*I*’ and ‘*we*’ in the Ancient Lives corpus

When the Zooniverse humanities corpora are compared against a sample of general Internet English drawn from the giant enTenTen13 corpus (Sketch Engine n.d.) the word *I* emerges as the highest-ranking keyword. While this is not necessarily surprising in itself, since it is common to refer to oneself in many informal and conversation-like settings, it is potentially significant insofar as *I* does not appear as a keyword in the ‘Dark Matter’ corpus when this is compared against the same general Internet corpus. Hamilton et al. (2017), in their study of loyalty in online Reddit communities, found that a relatively heavy use of *I* pronouns was associated with what they term ‘vagrant’ users, who drift from community to community rather than remaining loyal to just one. This ties in with our earlier remark, to the effect that the Zooniverse humanities communities tend to be less established in timescale than those accompanying the scientific projects. User behaviour may not yet have ‘settled down’ in terms of personal commitment and community norms. Whatever the case may be, the *I* phenomenon is deserving of further investigation.

In this section, the focus will be on just one of the humanities corpora – Ancient Lives – to allow us to home in on the context of a single subject environment and to explore how personal pronouns are used as the citizen scholars relate themselves to the subject matter and the people around them. This project asks members of the public to transcribe fragments of papyri from the Oxyrhynchus Papyri – a collection discovered in Egypt in the nineteenth century.

The highest frequency personal pronouns in the corresponding corpora are *I* (433 occurrences) and *we* (95 occurrences). WordSmith Tools 6.0 pattern view will be used to highlight the lexical environment of these two personal pronouns.

The ‘*I*’ environment

Figure 28.3 shows the pattern view for the lexical item *I* in the corpus. The central, empty column represents the node *I* and shows collocates in order of decreasing frequency. The word *that*, for example, is the most frequent L1 (= one place to the left) collocate and *am* is the most frequent R1 (= one place to the right) collocate. These correspond to the structures

TO	THE	THE	THAT		AM	TO	TO	TO
THE	OF	AND	BUT		HAVE	NOT	THE	THE
ON	AND	TO	AND		CAN	THE	THAT	THAT
IS	IN	OF	IF		THINK	IT	IT	OF
THIS	IS	IS	WHEN		WOULD	HAVE	OF	JUST
BUT	BE	THIS	SO		DON'T	THAT	IS	IS
IN	TO	MY	WHICH		CAN'T	BE	YOU	THIS
OF	VERY	THAT	IT		DO	SEE	HAVE	AND
GET	ON	HOW	DO		WAS	THINK	THIS	BUT
THAT	THAT	OR	HOW		GET	ON	MY	ON
NOT	FOR	IN	NOW		KNOW	LIKE	IF	IN
AND	FROM	NOT	GT		ALSO	AN	AND	IT
BE	IT	BUT	SYMBOL		HAVENT	JUST	AT	THERE
THEY	AT	AS	WHAT		HAD	READ	IN	HAVE
CAN	THANKS	THANKS	AM		FOUND	KNOW	ON	CAN
RIGHT	HELP	ANY	WAY		JUST	GET	AN	WRONG
OR	TEAM	VERY	STUFF		CLICK	ACROSS	LETTER	WITH
SOME	AS	HELPFUL	THING		SEE	SUGGEST	LITTLE	ABOUT
WITH	SO	WHAT	THIS		REALLY	TRANSCRIBE	TRANSCRIPTION	OUT
WHAT	THIS	HI	LETTERS		PERSONALLY	AT	TRANSCRIBING	SO

Figure 28.3 WordSmith tools pattern view for the node 'I' in the Ancient Lives corpus

that I and *I am*. The determiner *the* in L2, L3, L4 and L5 position draws focus to either the only noun group or the last noun group of a list before the pronoun *I* is produced. This structure is frequently used when the user is having problems with the site as can be seen in Text 1:

- 1 I came back to this project a few days ago, worked normally one evening long, and from the next day on . . . whenever I login and try to transcribe, I always get: 'Complete – You have viewed all fragments in the system. 'Is this a problem you are aware of or have I done something wrong?'

The *the* highlighted in pattern view is the beginning of the structure *the next day* and, in this case, provides temporal information about recent activities before the poster describes a problematic pop-up message.

The lexical items *thanks* and *helpful* stand out in the L2 and L3 columns as indicators of gratitude in structures such as *thanks for sharing!*, *thanks to all!!* and *it is very helpful* before the posters produce the pronoun typically to either offer support or ask for support in this environment:

- 2 The material in the Blog that I rediscovered has been/ and will be very helpful. And I found a couple of YouTube videos last evening with [name] that explained a lot too.

In Text 2 the poster shares information about a helpful blog before using *I* to highlight their active role in learning more about the subject matter. VOLCROWE (www.volcrowe.org) has shown that learning about the subject matter is an important motivator for site users.

I am is the most frequent 2-gram (= 2-word string) that contains *I*, with 43 occurrences.

The 2-gram is followed by 29 different R2 items ranging from *a* and *able* to *very* and *working*. Only five 3-grams appear more than once – 3 occurrences of *I am a*, 3 occurrences of *I am afraid*, 7 occurrences of *I am not*, 3 of *I am sure* and 2 of *I am very*. This R2

one something wrong?? Thanks and regards. I am having this problem as well. However, after I view a fi
ared the article with many people. Currently I am informing a lot of people about AL, especially those i
e transcription is fun & I as a history fan I am interested in taking part, but without even the chanc
g, sort of like grading us on a test. Right now I am just submitting work into some black hole with no ide
:0057 This is all my knowledge of the matter. I am looking forward to see more suggestions too. One otl
ar with greek characters and words. Hi there, I am Marsia Bealby, originally Greek (and my Ancient Gree
ters_without_transcription-Useful_or_not? I am new to this and very enthusiastic I came across this it
'copyrightable' As for the original text, itself, I am not a copyrights lawyer (nor any other kind of lawyer
nt looks like a piece of crumpled vellum and I am not sure of this identification. Hi, giorast. Thank you f
confusing variations of the letters (obviously I am not a linguist or philographer (SP)) Thanks Yes, it doe
nt? Ive been doing this for several days now. I am not an expert, but my understanding so far is everyth
ter I submit something, what is the process? I am not going to quit, but some introduction that explain:
1_Sigma_and_get_an_accent_mark_instead? I am not able to replicate your problem.... Is a browser prc
ir Transcribe page again! lower_case_y Uhm. I am not sure. I would suggest you post a message linked t
searchers can answer you in more depth! Hil I am part of the Ancient Lives team in Minnesota and we h
ir loop. I think the yellow ribbon symbol that I am seeing is just the last part of the omega written as a l
t server which appears to be part of the site. I am sorry if I made a mistake by making such a reference,
are it against other examples of letters which I am sure of. Itd be great if this could be a function. I seco
ele. Its great to have an expert user like you. I am sure we will benefit a lot from your contribution. Firsi
.richardcarrier.info/papyrus/script.html)but I am sure that there must be much more information about

Figure 28.4 Sample concordance of 'I am' in the Ancient Lives corpus. (The data have been selected to include the set of seven occurrences of 'I am not'.)

environment following *I am* suggests an audience that are coming together to learn something new – the volunteers are expressing gratitude for help, expressing their interest in the site and, often, marking their status as a beginner. Consider just the seven occurrences of *I am not* shown in Figure 28.4. Three are followed by a noun form as the posters stress they are *not a copyrights lawyer*, *not a linguist* and *not an expert*. Two form part of the expression *I am not sure* to mark doubt. The final two cases form part of the structures *I am not going to quit* and *I am not able to replicate your problem* as the posters stress determination in a new learning environment and try to support each other.

The final part of the *I* environment that we will draw examples from is based on the lexical item *to* (technically two lexical items with the same form) at the top of the R1 column. This highlights structures such as *which I replied to* where the poster comments on their activity interacting with other site users (typically *to* as preposition). It also includes structures where the site user is referring to things they have done, things they know and things they want such as *I forgot to post the link in my comment* (this will typically include *to* as the infinitive particle). 16 different forms sit between *I* and *to* in the dataset of 20 occurrences forming 3-grams such as *I know to*, *I seem to*, *I have to* and *I had to*. *I have to* is the most frequent *I * to* form with the following four occurrences:

- 3 do I have to measure when the zoom is not on
- 4 Till now I have to make many mouse clicks
- 5 How I have to figure out why the transcription tool is not working
- 6 I have to agree with these sentiments

It will come of no surprise that the form is being used to express a sense of obligation. Three of these (3, 4 and 5) are using *have to* to refer to what is needed to operate the software. In Text 6 the site user is referring to a shared sense of frustration with other

TO	THE	HI	AND	WE	ARE	BEEN	TO	TO
OF	OF	THE	IF		HAVE	THE	OF	ON
AND	YOUR	THAT	SO		CAN	TO	ON	IT
FOR	ON	AND	BUT		DO	BE	THE	THE
OUR	PROBLEMS	WITH	TRANSCRIBE		WILL	AWARE	ASKED	PERFORMANCE
WORK	IT	TRANSCRIPT	HOW		HOPE	THAT	OUR	YOU
FRAGMENT	AT	AM	UNFORTUNATELY		PLAN	NOT	HAVE	THESE
GREEK	ARE	ABOUT	YES		CANNOT	TRYING	USER	SOME
IN	FROM	OF	WHAT		WOULD	WORKING	ADDRESSING	OF
	IS	LIGHTBOX	CAN		COULD	ON	ABLE	FIXING
	IN		MOMENT		CANT	CURRENTLY	MORE	FIX
			THAT		MAY	ALL	EXPECT	AN
						ADDRESS		FROM
						NEED		NOW
						IDENTIFY		LIKE
						HAVE		LETTER

Figure 28.5 WordSmith Tools pattern view for the node 'we' in the Ancient Lives corpus

users at a feeling they are not sure if they are transcribing correctly and a perceived lack of support from page leaders. Across all 20 occurrences two main themes come through – the site users are sharing information about how to use the site and they are expressing a desire to help and get involved. The two cases that do not map on to these themes are Texts 7 and 8:

- 7 I had to study Ancient Greek in high school
 8 But I appeal to those who negotiate them

These posts are quite unusual in the way they reach out beyond direct references to transcribing data. In Text 7 the poster is explaining their knowledge of ancient Greek as they give information about their own background and offer suggestions for how leaders can better promote the site. In Text 8 the poster is appealing for anyone in an appropriate position to campaign for more open-access publishing in academia.

The 'we' environment

By way of comparison with Figure 28.3, Figure 28.5 shows the pattern view for the lexical item *we*. Similar to the *I* environment, the *we* environment as shown here suggests an atmosphere where the volunteers are very much focused on the task. The words *fragment*, *Greek* and *transcript* are visible as collocates of *we* on the left side of figure three and *working*, *addressing* and *performance* are on the right. At a glance, the R1 collocates suggest that *we* often refers to Zooniverse staff addressing problems and complaints – indeed *we are working on fixing that issue* and *we are working on improving the functions* are two of the text fragments that comprise the data.

The word *the* is the highest frequency item in the L3 position but, in this dataset, one must remember the numbers are very small. There are only 95 occurrences of *we*, and the item *the* in the L3 column refers to just three text fragments:

- 9 We are aware of the problem and we are trying to fix it. Thanks for your patience.
 10 but one gets the feeling that we may need to be very patient. I agree with you -
 11 sigma Σ on the keyboard because we hope put fragments like the Deverni papyri from

io could not spend our lifetime that way. We are all indebted to those who publish this stuff, right bac
n_link_doesnt_work_anymore Cheer up. We are all suffering. See way below your post: Trouble with
. Transcribe it will allow me to transcribe. We are aware of the problem and we are trying to fix it. Tha
npossible_by_dr_Strangelove Hi, giorast. We are aware of this problem and are trying to fix it. tech_r
n only rotate pieces in your Lightbox. But we are aware of the problem and working on that. Thanks fo
n only rotate pieces in your Lightbox. But we are aware of the problem and working on that. Thanks fo
re at least half of my papyri | cant rotate. We are currently working on putting more images into the s
a Microsoft Windows 7. Hi martharow18, we are currently experiencing some problems with Transcrit
: allowing the system to map it. The tools we are developing on an ongoing basis to study these fragm
: correct transcription so we can see how we are doing, sort of like grading us on a test. Right now I an
rtunately caused a series of broken links. We are fixing these issues one by one, in conjunction with th
hael Thanks for your comments Raphael. We are in fact now developing a way for users to see the fini
res, a Coptic part which is a bit tricky and we are still testing and improving it). There have been many
each published fragment. At the moment we are testing users classifications of published papyri and c
scribe. We are aware of the problem and we are trying to fix it. Thanks for your patience.
lives.org/help/discussions/DAL10030lo). We are trying to fix the problem. iPad Karen Hi Karen, We di
help revise and improve earlier editions. We are trying to get a consensus transcription from users' tr
rk spreading should do the trick :). Hello, we are two students from Greece and we go to Junior High S
ve have thought about it. At the moment we are very busy, analyzing the data that you have sent and
anything_to_transcribe_Hi bumblebee2, we are working on fixing that issue. thanks Marco.
cript. Just a brief note to inform you that we are working on improving the functions of the Match But
riencing some problems with Transcribe. We are working on fixing them. 1 do not think that issue def

Figure 28.6 Full concordance of 'we are' in the Ancient Lives corpus

As discussed earlier, the *the* * * *we* pattern gives us a glimpse of the last noun form being used before the writer refers to the group represented by *we*. It can be seen from examples 9–10 that the three noun forms are *the problem*, *the feeling* and *the keyboard*. Two out of the three examples are written by Ancient Lives staff (e.g. project managers or forum administrators); in both of these cases *we* refers to the Ancient Lives team as they assure the volunteers they are trying to fix a problem and explain details about the forms of the Greek letter sigma getting used in the period and which ones are available on the Ancient Lives keyboard. In Text 10 the *we* refers to the community of citizen scientists/scholars, arguably to people new to the project, as a new volunteer marks themselves as such and expresses the need for patience as they learn the basic skills needed.

Figure 28.6 shows the full concordance of *we are* in the Ancient Lives corpus. One striking difference between this and the *I am* data discussed earlier is the proportion of posts (18 out of 22 = 82 percent) by Ancient Lives staff. This suggests an atmosphere where the leaders consider themselves part of a group, but the citizen scientists/scholars are less inclined to refer to themselves as a group. In fact, *we are two students from Greece* refers to a group of two, so only three posts – *we are all indebted to those who publish this stuff*, *we are all suffering* and *so we can see how we are doing* – appear to construct the volunteer community as a group.

To recap, this project involves citizen scientists/scholars learning how to transcribe ancient papyri so that scientists have access to a machine-readable set of transcriptions. The analysis of the *I* environment gives us a glimpse as to how these volunteers see themselves as individuals. There are clear references to problems and difficulties with the software, but also that the site users are keen to stress their enthusiasm as they join the community. There

is also a clear sense that they are not experts and make that point clearly to other site users. The *we* environment also brings up references to the data and the act of transcription, but these data include a large proportion of posts by the organising team. This suggests that the ‘citizens’ in this particular project do not yet see themselves as a coherent group or a fully fledged community of practice.

Conclusions and future directions

This chapter sits at the convergence of two strands of the digital humanities, first through its methodology – corpus linguistics – and second through the online crowdsourcing activity exemplified by CS. Arguably, until very recently, this latter strand has been relatively neglected in ‘internalist’ accounts of the digital humanities – in spite of earlier crowdsourcing projects such as the ‘Transcribe Bentham’ initiative (Terras 2011), which sits clearly within the digital humanities field. It is the contention of the present authors that the crowdsourcing methodology offers significant new potential to the digital humanities, principally by bringing in new actors from outside the academy. Thus, we are pleased to see that crowdsourcing (via Zooniverse) has become a regular workshop theme at the annual Digital Humanities at Oxford Summer School (DHOxSS 2019).

Thus far, the humanities projects under the Zooniverse umbrella have taken ‘citizen science’ as their model, with the role of the general public being confined to the transcription and identification of predetermined objects of study. As such, individual identities, personal histories and value systems are not directly called into play. Under the heading of the public humanities, however, there is scope for broader and more personalised involvement in questions related to language and society, for instance, the crowdsourcing of linguistic judgements of correctness and acceptability. This would add an extra spin to the old debate between prescriptive and descriptive attitudes towards the use of language.

Central to the further development of the ‘citizen science’ model of the public humanities are questions of *community* and *language*. The Zooniverse forums are not an Internet free-for-all like Twitter or below-the-line commenting on press articles. They require a degree of commitment, collaboration and a certain learning curve prior to full participation in the research activities. At any given time, different individuals will have a different status within the community and be at a different stage in their personal trajectories. These different statuses and stages, and the changing relationships between participants, will – so this chapter has attempted to demonstrate – be marked by specific language features. For instance:

- The recurring semantic sequence ‘sorry for my English’ (and close variants) has been identified as a strategy adopted by non-native speakers of English to negotiate membership of the community. It is more broadly indicative of a novice/expert distinction that also manifests itself in other ways in the Zooniverse forums.
- These distinctions include the use of **NEGATIVE** + *or* in the humanities forums in order to de-emphasise one’s credentials. (By contrast, *or* does not seem to be used in this way in the scientific forums – quite the opposite in some cases.)
- This finding is reinforced by the detailed analysis of the *I* environment in Ancient Lives, though the analysis goes on to show that *I* is also used in other sequences that express community solidarity or frustration.
- The fact that, in the Ancient Lives forums, the pronoun *we* is used preponderantly by the site managers suggests that there might be different perceptions of the community.

At the current stage of research, these findings represent only a partial or provisional account of the language of the Zooniverse community. Findings based on one project corpus needed to be tested against the others, and a more thorough-going and systematic comparison needs to be undertaken between the humanities corpora and the natural science corpora to determine whether there are any significant disciplinary differences in CMC discourse and what these represent in terms of the self-perception of the respective communities. In particular, the wider analysis of a range of interpersonal and diachronic phenomena remains work in progress.

In time, these lines of research are likely to be of practical benefit to the organisers of CS and citizen humanities projects. Citizen scientists themselves are coming to realise – arguably somewhat belatedly – the importance of ‘language use’ in motivating or demotivating participants (see Eitzel et al. 2017). Our team members are working closely with colleagues in computing science and economics to explore the question of what language choices influence the success of a CS project. Issues such as the movement of individuals in and out of groups, power, gender, the effects of charismatic and/or overbearing individuals and various other aspects of how site users see themselves come up in different ways over and over again in our discussions; such discussions will inevitably intersect with the concerns of other researchers interested in different online communities.

Further reading

- 1 Hadikin, G. (2015). Lexical selection and the evolution of language units. *Open Linguistics* 1(1): 458–466.
- 2 Hunston, S. (2008). Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics* 13(3): 271–295.
- 3 Toerpe, K. (2013). The rise of citizen science. *Futurist* 47(4): 25–30.
- 4 Zooniverse.org (2017). *Welcome to the Zooniverse: People-powered research*. Available at: www.zooniverse.org

References

- Aarts, J. (1991). Intuition-based and observation-based grammars. In K. Aijmer and B. Altenberg (eds.), *English corpus linguistics: Studies in honour of Jan Svartvik*. London and New York: Longman, pp. 44–62.
- Allington, D., Brouillette, S. and Golumbia, D. (2016). Neoliberal tools (and Archives): A political history of digital humanities. *Los Angeles Review of Books* [online]. Available at: <https://lareviewofbooks.org/article/neoliberal-tools-archives-political-history-digital-humanities/>.
- Anthony, L. (2006). Developing a freeware, multiplatform corpus analysis toolkit for the technical writing classroom. *IEEE Transactions on Professional Communication* 49(3): 729–737.
- Anthony, L. (2016). *Laurence Anthony's website*. Available at: www.laurenceanthony.net/software.html.
- Brain Bank North West. (2015). *Citizen science: The power of the crowd*. Available at: <https://the-brainbank.scienceblog.com/2015/08/23/citizen-science-the-power-of-the-crowd/>.
- Bruthiaux, P. (2003). Squaring the circles: Issues in modeling English worldwide. *International Journal of Applied Linguistics* 13(2): 159–178.
- Crystal, D. (2006). *Language and the internet*. Cambridge: Cambridge University Press.
- Davies, A. (2003). *The native speaker: Myth and reality*. Clevedon: Multilingual Matters.
- DHOxSS. (2019). Crowdsourced research in the humanities (Theory and Practice). *Digital humanities at Oxford summer school*. Available at: www.dhoxss.net/crowd-sourced-research.

- Eitzel, M.V., et al. (2017). Citizen science terminology matters: Exploring key terms. *Citizen Science: Theory and Practice* 2(1): 1–20. Available at: <https://doi.org/10.5334/cstp.96>.
- Enteen, J. (2010). *Virtual English: Internet use, language, and global subjects*. London: Routledge.
- Firth, A. (2009). The lingua franca factor. *Intercultural Pragmatics* 6(2): 147–170.
- Gledhill, C. (2000). *Collocations in science writing*. Tübingen: Gunter Narr.
- Groom, N. (2017). *Locating semantic sequences in theory and practice*. Paper presented at Symposium on Corpus Approaches to Lexicogrammar, Edge Hill University, 10 June. Available at: www.edgehill.ac.uk/english/files/2017/06/LxGr2017.Groom_.pdf.
- Hacking, I. (1983). *Representing and intervening*. Cambridge: Cambridge University Press.
- Hadikin, G. (2015). Lexical selection and the evolution of language units. *Open Linguistics* 1(1): 458–466.
- Hadikin, G. (2017). *Lexical selection in citizen science: The selection of 'we are' in two online environments*. Manuscript submitted for publication.
- Hamilton, W., Zhang, J., Danescu-Niculescu-Mizil, C., Jurafsky, D. and Leskovec, J. (2017). *Loyalty in online communities*. Available at: [arXiv:1703.03386v3](https://arxiv.org/abs/1703.03386v3).
- Herring, S. (1996). *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*. Amsterdam: John Benjamins.
- Holliday, A. (2006). Native-speakerism. *ELT Journal* 60(4): 385–387.
- Humanities Blast. (2011). *The digital humanities manifesto 2.0*. Available at: www.humanitiesblast.com/manifesto/Manifesto_V2.pdf.
- Hunston, S. (2008). Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics* 13(3): 271–295.
- Illingworth, S. and Allen, G. (2016). *Effective science communication: A practical guide to surviving as a scientist*. Bristol: IOP Publishing.
- Kachru, B.B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In R. Quirk and H. Widdowson (eds.), *English in the world: Teaching and learning the language and literatures*. Cambridge: Cambridge University Press, pp. 11–30.
- Long, M.H. (1983). Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics* 4(2): 126–141.
- Luzón, M.J. (2013). Public communication of science in blogs: Recontextualizing scientific discourse for a diversified audience. *Written Communication* 30(4): 428–457.
- Mahlberg, M. (2007). Clusters, key clusters and local textual functions in Dickens. *Corpora* 2(1): 1–31.
- Manzoor, A. (2016). Online communities of practice and web 2.0. In A. Kok and H. Lee (eds.), *Cultural, behavioral, and social considerations in electronic collaboration*. Hershey: IGI Global, pp. 265–292.
- Martocchia, M. (2012). *Chasing solar storms on a wing and a game*. Available at: www.gamification.co/2012/04/19/chasing-solar-storms-on-a-wing-and-a-game/.
- Mercer, N. (2000). *Words & minds*. London: Routledge.
- Rampton, M.B.H. (1990). Displacing the 'native speaker': Expertise, affiliation, and inheritance. *ELT Journal* 44(2): 97–101.
- Rosenbloom, P.S. (2012). Towards a conceptual framework for the digital humanities. *Digital Humanities Quarterly* 6(2). Available at: www.digitalhumanities.org/dhq/vol/6/2/000127/000127.html.
- Saraceni, M. (2010). *The relocation of English: Shifting paradigms in a global era*. Basingstoke: Palgrave.
- Schildhauer, P. (2017). A diachronic approach to web-based genres: The case of the personal weblog. *Linguistik Online* 80(1): 95–124.
- Schreibman, S., Siemens, R. and Unsworth, J. (eds.). (2004). *A companion to digital humanities*. Oxford: Blackwell. Available at: www.digitalhumanities.org/companion/.
- Scott, M. (1999). *WordSmith tools manual* (version 3). Oxford: Oxford University Press.
- Scott, M. (2017). *WordSmith tools*. Available at: www.lexically.net/wordsmith/.

- Sinclair, J. (1996). The search for units of meaning. *Textus* 9: 75–106.
- Sinclair, J. (2005). Corpus and text – basic principles. In M. Wynne (ed.), *Developing linguistic corpora: A guide to good practice*. Oxford: Oxbow Books, pp. 1–16. Available at: <http://ota.ox.ac.uk/documents/creating/dlc/>.
- Sketch Engine, (n.d.). Available at: www.Sketch Engine.co.uk/.
- Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
- Terras, M. (2010). A decade in digital humanities. *Journal of Siberian Federal University: Humanities & Social Sciences* 7(29): 1637–1650.
- Terras, M. (2011). Present, not voting: Digital humanities in the Panopticon. *Literary and Linguistic Computing* 26(3): 257–269.
- Toerpe, K. (2013). The rise of citizen science. *Futurist* 47(4): 25–30.
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.
- Varonis, E.M. and Gass, S.M. (1985). Nonnative/nonnative conversations: A model for negotiation of meaning. *Applied Linguistics* 6(1): 71–90.
- Vettorel, P. (2014). *English as a Lingua Franca in wider networking: Blogging practices*. Berlin: Walter De Gruyter.
- Virtanen, T. (2017). Adaptability in online consumer reviews: Exploring genre dynamics and interactional choices. *Journal of Pragmatics* 116: 77–90.
- Virtanen, T., Herring, C. and Stein, D. (2013). *Pragmatics of computer-mediated communication*. Berlin: De Gruyter Mouton.
- Wenger, E., McDermott, R. and Snyder, W.M. (2002). *Cultivating communities of practice: A guide to managing knowledge*. Boston, MA: Harvard Business School Press.
- Wenger-Trayner, E. and Wenger-Trayner, B. (2015). *Introduction to communities of practice*. Available at: <http://wenger-trayner.com/introduction-to-communities-of-practice/>.
- Williams, J. (in preparation). ‘Do you think you could dumb it down a notch?’ Using lexis to explore tensions in online ‘citizen science’ discussion.
- Winder, J. (2012). *Wishing the Zooniverse a happy 5th birthday!* Available at: www.universetoday.com/96145/wishing-the-zooniverse-a-happy-5th-birthday/.
- Winters, J. (2017). Tackling complexity in humanities big data: From parliamentary proceedings to the archived web. *Studies in Variation, Contacts and Change in English*: 19. Available at: www.helsinki.fi/varieng/series/volumes/19/winters/.
- Wooldridge, R. (2004). Lexicography. In S. Schreibman, R. Siemens and J. Unsworth (eds.), *A companion to digital humanities*. Oxford: Blackwell. Available at: www.digitalhumanities.org/companion/.
- Yano, Y. (2009). The future of English: Beyond the Kachruvian three circle model? In K. Murata and J. Jenkins (eds.), *Global Englishes in Asian contexts: Current and future debates*. Basingstoke: Palgrave, pp. 208–225.
- Zhang, Y. and Vásquez, C. (2014). Hotels’ responses to online reviews: Managing consumer dissatisfaction. *Discourse, Context & Media* 6: 54–64.
- Zooniverse.org. (2017). *Welcome to the zooniverse: People-powered research*. Available at: www.zooniverse.org.

English language and digital cultural heritage

Lorna M. Hughes, Agiatis Benardou and Ann Gow¹

Introduction

The chapter will address the ‘datafication’ of cultural heritage content (tangible and intangible) and the use and reuse, and potential, of this content for humanities research. This process is highly collaborative, drawing into partnership researchers from scientific and other humanities disciplines, computational and technical fields, as well as cultural heritage organisations, it also relies on underlying technical and enabled research infrastructures. There are significant advantages to digitising and creating online access to cultural heritage content. The greatest of these advantages is the creation of new resources for scholarship via the development of an emerging ‘digital commons’ for language research; one that is the basis for future production as new research emerges and new ways of working with the content are perfected. Through this, new communities of practice are fostered. Ideally, this process will also make visible intellectual content of archives and heritage collections that has hitherto been hidden or inaccessible.

Background

The past 20 years has seen an enormous investment in the digitisation of Europe’s cultural heritage. Each year, the ENUMERATE organisation carries out a study of the extent of Europe’s digital cultural heritage. This survey is commissioned by *Europeana*, Europe’s digital platform for cultural heritage, which aggregates digital collections from more than 3000 institutions across Europe, letting users explore Europe’s collective cultural and scientific heritage from prehistory to the modern day. In the 2017 survey, ENUMERATE assessed that 22 percent of the collections of European museums, archives and libraries has been digitised, a total of over 300 million items (Europeana 2017). It will take over 30 years to digitise the what remains, much of which is twentieth- and twenty-first-century material, composed of moving image material, audio or born-digital.² A growing number of organisations have developed digital strategies to make their content accessible: of the memory organisations that responded to the ENUMERATE survey, 84 percent have a digital collection (Ibid.).

The expansion of digital cultural heritage from library, archive, museum and university collections has created tremendous opportunities for research. The ‘digital turn’ has also enabled innovative approaches to representing spatial and temporal aspects of primary source materials. However, challenges still exist in accessing and using digital heritage, including access to the content, availability and use of digital tools and methods and the publication of research outputs in ways that are recognised by peers. Furthermore, there is still very little understanding of how this content – much of it including significant sources for language research – is used in research. Among humanities researchers, there still exists limited awareness of how to integrate this material into the research life cycle.

There has, of course, been a profusion of digital humanities initiatives, and the development of humanities approaches to working with digital collections at scale has paralleled this growth in digital cultural heritage. Many early digital humanities projects explored ways to answer humanities research questions by working with digital data, including the Oxford English Dictionary (Bradley et al. 2016); the Thesaurus Linguae Graecae (TLG); the Perseus collection of Greek and Roman materials; and corpora including the LOB corpus and the British National Corpus (Rayson et al. 2017: 27–49). The use of the concept of ‘digital humanities’ to describe this activity has been current since the publication of the *Blackwell Companion to the Digital Humanities* in 2004 first deployed the term. Since then, the concept has become a ‘big tent’ to describe approaches by humanities, arts and social science researchers to explore ways that experimentation with (at first computational, then digital) tools, techniques, interventions and media can alter the production, analysis and dissemination of knowledge. Exciting work flourished in what Greg Crane at the University of Leipzig has called *e-Wissenschaft*,³ reflecting digital humanities as a new intellectual practice entirely divergent from print-based scholarly approaches (Crane 2009). The rapid growth of the field has led to an increasing popular awareness of digital humanities as what Matt Kirschenbaum (2012) called a ‘free floating signifier’: a label for self-identification, which is beguilingly inclusive when scholars are all, to a degree, digital – using digital data for research by default, or at the very least, electronic catalogues that point to analogue resources (Ell and Hughes 2013). Most scholars now rely on a steady consumption of digital source and catalogue materials for scholarship and pedagogy, mostly created through large-scale digitisation initiatives in universities, libraries, museums and archives, as well as by commercial entities. But the impact of digital in scholarship extends beyond consumption and use of data: many scholars now also create and manage these resources and are engaged in digitally enabled scholarly communication: whether through online publishing or social media engagement, digital dissemination methods are essential for sharing and challenging research results. For many, this digital disruption in the scholarly life cycle represents a sea change in research practice within the recognisable lifetime of a career.

There are almost as many definitions of digital humanities as there are digital humanities (DH) centres.⁴ DH Centres can be initiatives, labs, collectives and networks, ranging from small labs and virtual support centres to large-scale physical institutes, with emerging technologies and support, to academic departments which award degrees. Transnational infrastructure initiatives such as DARIAH (the Digital Research Infrastructure for the Arts and Humanities, www.dariah.eu) can also be seen. However, DH can be broadly defined as a way of carrying out research in the humanities that combines several things: using digital content, applying digital methods for the analysis and interpretation of this content and using digital tools for specific scholarly tasks. This research is communicated to the widest possible audience using traditional and non-traditional publishing methods. The benefits of DH approaches are threefold: they facilitate and enhance existing research, making some

research processes easier via the use of computational tools and methods, and they enable new research that would be impossible to undertake without digital resources and methods, and they inspire new research questions that are driven by insights only achievable using new tools and methods (Hughes 2011).

These approaches typically rely on collaboration: with researchers from scientific and other humanities disciplines and computational and technical fields, as well as cultural heritage organisations. Like any research activity based within the higher education sector, this relies on underlying technical infrastructures. Another important collaboration is with the audiences and consumers of this research, as user-led design, and participation via mechanisms like crowdsourcing currently inform the development of many initiatives. This characterisation of digital humanities as a space (or series of processes) that incorporates *content, tools and methods*, evokes a working environment that can be mapped to the analogue scholarly environment: it brings together data as the primary sources for scholarship: tools for working with the sources and expertise in the form of digital research methods. It also acknowledges the ‘positivist, objectifying, task-oriented and practical approach’ described by Hall (2012). However, as Liu (2012) has noted, digital humanities is also a means of cultural critique. As such, it is one where the maker and theorist perspectives are not incompatible: through developing and building digital projects in the humanities – *through practice* – cultural and critical analysis can be affected, questioning many of the assumptions on which digital resources are built and communicated and developing a better framework for understanding the ways that working with digital content and digital infrastructure is transforming our consumption and production of knowledge.

Critical issues and topics of digitally enabled scholarship and the use of digital methods

Digitally enabled scholarship in the humanities is far more than just discovering and appraising the suitability of digital sources for research and scholarship: it also involves consideration of the information needs, practices and methods of scholars as they pursue their research. The EU-funded project, *Europeana Cloud: Unlocking Europe's Research via The Cloud* addressed this issue in depth, specifically in a work package on user requirements, which carried out a review of review of some relevant research projects and studies that build a clearer picture of the role of digital tools, research processes and content that are used in the humanities. The findings of this collaborative project are discussed below and the following section is based on the final project report.^{5,6}

The rising impact of digital technology on scholarly research

In recent years, information behaviour research has explored the disruption in the humanities research process that has been made possible by access to digital resources, tools and services and the methodological promise and concerns related to the emergence of DH. A key area of enquiry has been how scholarly practice has been affected and to what extent the digital mediation of accessing, managing and using scholarly resources boils down to a standard set of granular micro-activities, or ‘scholarly primitives’.⁷ This has led to the emergence of new literature focusing on and exploring changes that have occurred, with the aim of recording and understanding new forms and dimensions of the research process. Apart from perceptions traced in the information behaviour literature since the 1950s, new approaches have emerged inside the context of digital scholarly information practice,

characterising current research in this field. A recent example can be seen in the work of Donald Case, who has suggested that the last 30 years have witnessed a shift from studying ‘the use of channels and sources’ to ‘encountering and seeking of information and the interpretation of meaning from that information’, resulting in an expanding ‘scope of investigation to view the use and user preferences for digital tools and services’(Case 2012: 153–154).

Since the beginning of the century, a growing range of research publications and projects have addressed the impact of the inception of digital technologies on research practice in the humanities (Houghton et al. 2004). One of the most interesting aspects of the impact of the digital age on research practices of humanists is their well-attested reluctance to adopt digital methods in the research process.⁸ Nevertheless, the digital age has resulted in notable changes in the research practices of the growing community of researchers who do use digital tools and content. As William A. Wulf (1998) points out: ‘the Humanities offer a new opportunity to explore how information technology can be employed in fundamentally different ways that will provide fresh insights and enrich research in other applications’ (Ibid 1998). While in the context of broader changes, such as the adoption of digital infrastructures in the hard and natural sciences, the use of digital technology by the humanities has significant differences regarding both researcher information behaviour and attitudes. While John Houghton, Colin Steele and Margaret Henty note the hesitation of humanists to adopt digital sources and practices, merely acknowledging that ‘they have not ignored the technology’ but insisting on the existence of a disciplinary divide with the sciences in terms of communication, collaboration, networking and in the form of sources used (Houghton et al. 2004: 51), other studies describe humanities scholars in particular as increasingly enthusiastic users of digital resources (Bulger et al. 2011: 10).

The growing amount of literature, research projects and digital infrastructures for humanities research mirrors the rise of the impact of digital technology on scholarly research practices, including publication and collaboration. As technology evolves, digital research practices and related literature will closely follow the extent of technology adoption. A point made by Griffiths et al. (2006: 5) perfectly captures this constant development: whereas humanities scholars were initially portrayed as hesitant users of digital technologies, ‘younger generations of scholars rely so heavily upon electronic resources that content available only in print has little visibility’. The reason underlying this phenomenon may be the easy colonisation and efficient use of the digital environment by younger scholars, who perceive it as a comfortable domain for conducting their own research. Unlike many researchers of the previous generation, younger scholars approach their research digitally since its beginning – this, in fact, is their ‘standard’ or ‘conventional’ way of conducting research. This generational change has a significant impact on the shift of attention, and value, from analogue to digital resources; as noted by the 2006 ITHAKA Report (Rutner and Schonfeld 2012), historians justifiably ‘worry that for undergraduates “if it isn’t electronic, it doesn’t exist”’ (Ibid: 12).

Digital content use by scholars

As researchers and scholarly resources constitute two core elements of the evolving scholarly process, the *Europeana Cloud* project sought to explore in detail the diverse dimensions defining and affecting how these two elements interact by looking at the active debate on the issue in scholarly literature. Nowadays, the increasing domination of digital content in scholarly practice gives rise to various questions regarding

the ways in which researchers identify, access and use content. For instance: How does digital content change research practices? How do researchers from different disciplines interact with different types of objects? Does this have an impact on the outcome of a researcher's work?

The topic of scholarly use of digital content figures extensively in information-seeking behaviour literature. The multidisciplinary study by Houghton, Steele and Henty serves as an introduction to the changing research practices of all disciplines and on the significance attributed to digital or analogue sources by researchers, including those in the arts and humanities and the social sciences (Houghton et al. 2004). Another study, which drew on the United Kingdom Research Assessment Exercise (RAE) 2001 for its arguments on the nature and range of required research material, nature of access to this material and its location is 'Researchers' Use of Libraries and Other Information Sources: Current Patterns and Future Trends' by Education for Change LTD, University of Brighton and The Research Partnership (Education for Change Ltd, SIRU 2002).

Yet another study, conducted by the University of Washington in 2005, similarly approached scholarly content use in the context of information behaviour of scholars (Bass et al. 2005). The study, based on an examination of information behaviour of researchers in the disciplines of humanities and social sciences seen as a unified field, focused on similarities rather than differences regarding information approaches and methods used by scholars. In contrast, a survey on the use of e-resources for research in the humanities and social sciences, published by the British Academy in 2005, seeks to throw light on the differences in research approaches advanced, respectively, by the humanities and the social sciences as they engage with digital resources (Jones and Bennett 2005).

As early as 2003, Meho and Tibbo had stressed the advantages offered by the Internet to social scientists in reaching digital information easily; however, they did note that, at that time, 'the problem is that only a small fraction of available information is digitised and made accessible to researchers' (Meho and Tibbo 2003: 586). Eight years later, the Research Information Network Report stressed the problem of assessing digital resources (Bulger et al. 2011: 72). This time gap allowed for an observation of the various issues arising in the literature in different chronological periods mirroring thus a discussion accompanying digital research practices along its development, an observation that seems to agree with Claire Warwick's et al. paper on 'Library and Information Resources and Users of Digital Resources in the Humanities' published in 2008.

A systematically theorised approach, based on the development of a formal conceptual model of scholarly information activity, has been introduced by Agiatis Benardou et al. (2013) in their paper on working practices, digital content and its usability and privacy issues of the research community of Holocaust researchers as a study group.

What about tools?

Unsworth (2003) considered many of the questions regarding the development and use of digital tools. He noted the potential of using web services to avoid researchers the hassle of installing software and dealing with hardware issues by requiring them to only use a web browser. Initiatives to explore this idea included the DHW project (Bia 2005), the Salami Project,⁹ the Monk Project¹⁰ and SEASR – The Software Environment for the Advancement of Scholarly Research.¹¹ The Digital Research Tools wiki (DiRT [developed from BambooDiRT]) is a community oriented towards an inventory of digital tools, providing updated information about how the field is developing.

The concept of digital tools and infrastructure as a significant area for research into the use of digital content for research was discussed as part of *The Cologne Dialogue on Digital Humanities 2012* (Thaller 2012). Here, Anderson et al. (2012: 147) proposed the idea of:

digital ecosystems [. . .] derived from communities rather than technologies. As open systems, digital research ecosystems will rely on communities and community involvement in a scenario where anyone can participate. The digital ecosystem is not for the specialist few but is instead about increased participation, sharing and building a social network of people, things, content and so on.

Future directions: bringing content, methods and tools together through the potential of research infrastructures

The *Europeana Cloud* Project also noted that one of the most visible manifestations of the strong presence of the digital in the humanities research process is the long list of research initiatives which have emerged during the last decade successfully exploring this area. More particularly, a significant number of research infrastructures in the humanities and social sciences have directed their efforts towards identifying current research patterns, developing tools and enlarging the pool of digital resources available to researchers. Among user-centric infrastructures, notable examples include the Digital Research Infrastructure for the Arts and Humanities (DARIAH-EU) and its preparatory project, Preparing DARIAH. In the context of DARIAH, various research results have been developed such as TextGrid and Isidore (Virtual Research Environment for the Humanities 2013; Isidore 2013). TextGrid constitutes a virtual research environment which enables researchers in the text-based humanities to create, analyse, edit, store and publish their research data in one (user) environment, while Isidore is an online system that collects, enriches and provides a unified data access for the digital humanities and social sciences. Another initiative has developed EASY, an electronic archiving system able to collect, archive and offer access to thousands of datasets in the humanities, social sciences and other disciplines (EASY 2013). Other European initiatives addressing research infrastructure include the European Holocaust Research Infrastructure (EHRI), the COnnecting REpositories (CORE) project, the COllaborative European Digital Archive Infrastructure (CENDARI) and the European Science Foundation Network for Digital Methods in the Arts and Humanities (NeDiMAH) (funded from 2011–2015).

Aiming to develop an open, collaborative research environment which will provide integrated online access to dispersed (archival) resources relating to the Holocaust, EHRI constitutes a major research project which takes under consideration various aspects of digital research to lay the foundation for building such an infrastructure. Furthermore, the CORE project, in conjunction with the ServiceCORE project, as well as the Digging into Connected Repositories (DiggiCORE) project, aim in general to facilitate free access to content stored across open-access repositories and may thus be considered as large aggregators of content. CENDARI, on the other hand, is a more specialised initiative which aims to integrate digital archives and resources for research on medieval and modern European history. Finally, NeDiMAH engaged in a series of activities and networking events that examined the practice of, and evidence for, digital research in the arts and humanities across Europe, centred around the concept of digital research methods. An output of the project is the NeDiMAH Methods Ontology (NeMO), developed through a collaboration with the Digital Curation Unit, Institute of Management and Information Systems, Athena RC (Hughes et al. 2015).

Focusing on digital resources, the Council of European Social Science Data Archives (CESSDA) constitutes the prime European infrastructure which contributes in locating, accessing and sharing data for the social sciences (CESSDA 2012). The Common Languages Resources and Technology Infrastructure (CLARIN), on the other hand, focuses more on linguistic resources.

Emerging research and key analysis: understanding the production and use of digital resources

Despite the promise of (and investment in) digital content, methods and tools and the infrastructures that support them, as noted earlier, there is a need for a critical framework for the use of digital cultural heritage in humanities research, and the following section describes a proposed approach to address this challenge.

Much of what is classified as ‘digital humanities’ is consumption: scholarly use of digital resources and born-digital material, working in what Roy Rosenzweig (2010) called a ‘culture of abundance’. While this presents new possibilities, it also raises significant challenges. As Tim Hitchcock (2013: 9–23) has described in his article ‘Confronting the Digital: Or How Academic History Writing Lost the Plot’, researchers often work with data that are extremely limited in their potential for reuse, as issues with optical character recognition (OCR), mark-up and description and provenance information may reduce the value of the digital material. The classic example is that of Google Books: visualisations of the data in Google Books is compellingly accessible through the easy availability of the Google Ngram Tool. However, this material is often used to support assertions that do not factor in the changes in historical language use, the incomplete nature of the resource and poor-quality OCR of the sources (Chumtong and Kaldewey 2017). The issue here is the ease of use and access supports the unquestioning use of a digital source. Similarly, limitations in metadata often means that the only way to engage with digital archives is through keyword searching, giving a ‘quick hit’ of a result but missing important contextual information that enables mindful engagement with archival content. For example, questions of scale of the original sources, their condition and their context are often hidden when using digital surrogates that often appear as disembodied objects. Ryan Cordell (2015), who has theorised a ‘network-author’ function in antebellum newspapers from work with digital newspaper archives, has noted that ‘most digital archives hide more than they reveal, as keyword searches require prior knowledge of the texts to be discovered and can lead to evidentiary excess’.

Another issue is the selection of content for digitisation. Due to limitations of funding, availability of analogue sources and copyright and licensing issues, many digital archives are not ‘complete’. The National Library of Wales received funding from the Welsh government and ERDF to digitise 1 million pages of Welsh newspapers from its holdings. These date from 1800–1919 in Welsh and English.¹² As a selection of the newspaper coverage of the time, this is an important resource for research and teaching: however, it is by no means a full representation of all newspapers of the period. From the earlier period, many of the newspaper runs held by the library are incomplete, and rights issues have meant that not all the newspapers held by the library are in the online collection. However, for the researcher, it’s too easy to assume that the resource is a complete representation of *all* Welsh newspapers, ever. Users (especially students) frequently turn to digital content by default when looking for historical sources, and as a free resource, this newspaper archive is very likely to be used. So, it needs to be clearly stated what the full scope of the project is as part of the collections description.

Creating environments for better use, reuse and linking of data

Another issue for researchers is the need for better analysis and linking of digital content, which is all too frequently locked in digital silos, unable to be openly reused and linked with other content. This is particularly appropriate to cultural heritage digital resources. Projects such as *Europeana* were developed to enable access to rich heritage resources across different memory institutions, but political and financial resourcing often means that the digital reflects the physical structures and access to collections is dictated by the institution (Department for Digital, Culture, Media & Sport 2018). Evaluation of users of digital content shows that scholars frequently have very simple questions or ideas that they want to test with data at the desktop, and they do not want the technology to be a barrier (Nichols and Clark 2015). However, creators of digital resources infrequently anticipate the use of tools for analysis and linking into the development of their content. Digital projects ideally should become open sandpits for experimentation, disassociating text and data from platform and delivery mechanisms, liberating digital resources for purposes unanticipated by the creators of digital resources.

For example, users frequently want to be able to use simple, pattern-finding tools, like Ngram, with a range of datasets to quickly test hypotheses through finding patterns – identifying ideas and concepts. While Ngram and related tools have their detractors, there is a compelling case for its use as a simple time-saving aid for researchers.¹³ Similarly, researchers frequently wish to publish and annotate data they find in relation to other materials. When linking and sharing social media content are straightforward, linking digital research sources related content is still a challenge – integrating these data seamlessly into publications and linking back to the original data are even more difficult (Robinson 2014). The *Europeana Research* initiative was established to develop research-led projects that utilised the Europeana ecosystem of collections and toolsets to address this issue, supporting projects that explored the creative use and re-use of collections, and the reports of the funded projects outline some of the challenges that arose.¹⁴

One of the huge advantages of digital access is making the invisible visible in archives. We lose this when we cannot easily reuse and re-purpose this content. There is a great deal of rhetoric about the ‘digital commons’ as the basis for future production – yet the practical implementation of these approaches is still a challenge. This is not to say the challenges are not recognised by the communities but that infrastructures, financial resourcing as well as technical challenges require complex solutions involving stakeholders from organisational governance through to end users.

Co-production

In many cases, digital archives and resources for scholarship, including literary and linguistic resources, have been created by memory organisations, where the praxis-driven development of digital heritage content is driven by access, relatively uninformed by the emerging critical and theoretical framework of digital humanities, and the emerging research agendas presented by the development of large-scale digital collections. For researchers in digital humanities to make effective use and reuse of digital content, the creation of content, collection building and curation should foreground processes of co-production with galleries, libraries, archives and museums. An example of this sort of co-production, or *co-creation*, can be seen in the development of the Jisc-funded mass-digitisation project, *Cymru1914.org: The Welsh Experience of The First World War*. This has created access to

digital sources from a variety of archives and special collections, with an underlying technical architecture supporting an open-source digital repository which is extensible: more content can be added over time, making this a 'living' digital archive. *Cymru1914.org* was a collaboration between the National Library of Wales, the Universities of Wales, partner archives, communities and researchers, bringing together research, curation and archive management to re-imagine the creation of digital content as a process of creative making, bringing together scholars and curators in ways that recast the role of each, driven by a shared goal of developing an integrated digital archive that was inclusive and enabled the reassessment of historical sources through its reuse for new purposes by a variety of audiences over time. The project presents a coherent, consolidated digital collection revealing the often-hidden history of the First World War as it impacted all aspects of Welsh life, language and culture for teaching, research and public engagement. The material has been reused in a number of ways: the OCR processed Welsh and English language content in the newspapers of the period was made accessible as a language corpus and has been used by the language research Centre Canolfan Bedwyr at Bangor University to study changes in the Welsh language during this period.¹⁵

Conclusion

A praxis-based critical engagement with digital content, tools and methods is the key to understanding the ways that the digital is affecting knowledge production: how the digital is drawing us into new collaborations, leading us to encounter new methods for engaging with content and assisting in developing new insights into our cultural heritage. A key challenge to humanities research is the use and reuse of the rich collection of digital primary sources and the data that they generate. Understanding information use by humanities researchers and the needs of scholars in the digital realm is a crucial aspect in embracing the opportunities of digital humanities: how can humanities research be carried out in Rosenzweig's digital 'culture of abundance'? A better understanding of the place of digital collections in the research lifecycle can also provide a critical framework for future development of digital collections by galleries, libraries, archives and museums (the so-called GLAM) sector. This perspective also provides evidence that can be used in the development of the research infrastructures that share this content and make it openly accessible for use and reuse.

While the focus has been to reflect on the challenges and successes of digital cultural heritage in English language research, the use of these resources also underpins research-led teaching. This is exemplified in the digitally enabled teaching at a UK Russell Group University, the University of Glasgow. The unique undergraduate degree of digital media and information studies delivered by the Information Studies Department grew from research into the digital in the English Language Department, and the synergy that exists between the two subject areas enables digitally enriched teaching experiences for the students accessing the courses offered through the degree programmes. Taking a collections-based approach to teaching and learning, the University of Glasgow's information studies undergraduate and master's degree, *Digital Media and Information Studies*, provides an interdisciplinary perspective on how the use of digital and analogue information and related technologies shapes our knowledge and understanding; provides new ways to view our past, present and future; and influences our economic, social and cultural life. Drawing on the disciplines of digital humanities and information science, digital media and information studies examines people's relationship with information and information and communication technology (ICT). It takes a critical perspective on the creation and use of digital media; how ICT

influences our analysis, interpretation and communication; shapes cultural heritage organisations; and the broader socioeconomic and cultural impact of the digital revolution. The interdisciplinary nature of the programme enables students from English language (as well as other arts and social science subjects) to share teaching events. For example, an honours course, Document Encoding which examines the relationship of textual sources with digital tools, is structured to engage students from various disciplinary backgrounds to develop shared perspectives of digital scholarship. In this case, the University of Glasgow's teaching must be research-led,¹⁶ forward-looking and adaptable to the needs of students worldwide. Embedding the issues of digital cultural heritage into the curriculum offers the opportunity to develop the infrastructure from a student and professional perspective. As with discussions about digital cultural heritage, the new is hinted at as the past is reflected on, following and shaping cultural trends to identify the local within a global practice.

As examined in our analysis of information seeking behaviours, identification and use of the digital resources we see scholarly activity embracing these resources and as the digital and artificial intelligence (AI) natives enter the faculty, we should reflect on the expectations and information literacy of the future. It is not always certain that familiarity with digital infrastructure will necessitate skills in recognising the quality and value of digital resources and so the academic community, through digital humanities units must ensure that digital enabled scholarship and methods are embedded in research and teaching activities.

Notes

- 1 Hughes and Gow authored the following sections: 'introduction'; 'background'; 'emerging research and key analysis: understanding the production and use of digital resources'; and 'conclusion'. Agiatis Benardou authored the sections: 'critical issues and topics of digitally enabled scholarship and the use of digital methods'; and 'future directions: bringing content, methods and tools together through the potential of research infrastructures'.

The authors would like to acknowledge the invaluable assistance of Sarah Gambell, University of Glasgow.

- 2 See the 2017 ENUMERATE – Survey Report on Digitisation in European Cultural Heritage Institutions: A summary of the 2017 results is also available at <https://pro.europeana.eu/page/digital-heritage-indicators>.
- 3 See www.digitalhumanities.org/dhq/vol/3/1/000023/000023.html
- 4 See <http://dhcenter.net.org/>
- 5 Europeana Cloud: Unlocking Europe's Research via The Cloud (pro.europeana.eu/project/europeana-cloud) coordinated by the Europeana Foundation, co-funded by the CIP-ICT Policy Support Programme: http://ec.europa.eu/ict_psp. CIP-ICT-PSP-2012-6 – Project number 325091. The project ran from 2013–2016. Agiatis Benardou was co-lead of the Work Package on user requirements, the findings of which are replicated here.
- 6 Europeana Cloud Deliverable 1.3 User Requirements Analysis and Case Studies Report. Content Strategy Report (Joint Deliverables 1.3 and 1.6: User Requirements Analysis and Case Studies Report (1.3) and Content Strategy Report (1.6)) (Angelis et al. 2015).
- 7 The term 'scholarly primitives' was first introduced by Unsworth (2000), who used this term in order to describe 'basic functions common to scholarly activity across disciplines', Unsworth, "Scholarly primitives: What methods do Humanities researchers have in common, and how might our tools reflect this?", *Symposium on Humanities Computing: Formal Methods, Experimental Practice*, King's College, London, 2000, <http://jefferson.village.virginia.edu/~jmu2m/Kings.5-00/primitives.html>, (Accessed: 19 June 2017).
- 8 See, for example, the perceptions of digital content and practices within history in Houghton, Steele and Henty, *Changing research practices in the digital information and communication environment*, PhD, Department of Education, Science and Training, 2003, p. 10.
- 9 For more information see: <http://ddmal.music.mcgill.ca/research/salami>
- 10 For more information see: <http://monkproject.org/>

- 11 For more information see: www.seasr.org/
- 12 Welsh Journals Online: <http://welshjournals.llgc.org.uk/>. A study into the use of Welsh Journals Online was written about in Hughes, L. M. (2012) Live and kicking: the impact and sustainability of digital collections in the humanities. In: C. Mills, M. Pidd, and E. Ward. eds. *Proceedings of the Digital Humanities Congress 2012*. HRI Online Publications: Sheffield, available from: www.hrionline.ac.uk/openbook/chapter/dhc-hughes (accessed 1 June 2017).
- 13 See, for example, Blanke, T. and Prescott, A. (2016) Dealing with big data. In: Griffin, G. and Hayler, M. (eds.) *Research Methods for Reading Digital Data in the Digital Humanities*. Series: Research methods for the arts and humanities. Edinburgh University Press: Edinburgh, pp. 184–205.
- 14 For reports of all projects funded see: <https://pro.europeana.eu/services/facilitating-innovation/grants-programme>
- 15 For more information see: www.bangor.ac.uk/canolfanbedwyr/
- 16 For more information see: www.teaching-matters-blog.ed.ac.uk/research-led-teaching-and-learning-in-search-of-meaning/

Further reading

Digital cultural heritage is a field where there are frequent new developments, and it can seem that the literature is constantly changing and difficult to keep up to date as a researcher. However, the key thing is for the scholar to be able to develop a critical framework for working with digital cultural heritage and using digital humanities approaches to working with this material. This will enable the scholar to understand the core principles underpinning the use of digital cultural heritage, and this will enable them to take an informed approach to selecting digital content to analyse and understanding how it has been created; evaluating the digital tools and methods that can be used to analyse this material and how they are supported and sustained over the long term; and the enabling infrastructures that support this sort of scholarship.

The following will help to cultivate this approach:

- 1 Hughes, L.M. (2004). *Digitizing collections: Strategic issues for the information manager*. London: Facet.
This was a very early guide to understanding how and why digital cultural heritage is created and the decision-making processes that underpin the way it is presented and preserved. It also introduces a key critical framework for digital heritage that still applies: understanding selection of content and the decisions about its digitisation; the use of metadata to represent content and encoding and encoding and enriching; the development of access and functionality, as well as the sustainability of content.
- 2 Prescott, A. and Hughes, L.M. (2018). Why do we digitize? The case for slow digitization. *Archive Journal* 6. Available at: www.archivejournal.net/essays/why-do-we-digitize-the-case-for-slow-digitization/
This article reviews some of the decisions underlying mass digitisation of historic manuscripts and argues for a greater awareness of the informational value of sources to inform digital capture and delivery.
- 3 Hughes, L., Constantopoulos, P. and Dallas, C. (2014). Digital methods in the humanities: Understanding and describing their use across the disciplines. In S. Schreibman and R. Siemens (eds.), *The new companion to digital humanities*. Oxford: Blackwells, pp. 150–170.
This chapter outlines work undertaken by the European Science Foundation Network for Digital Methods in the Arts and Humanities to map the use of digital content, tools and methods. It is a comprehensive attempt to formalise the practice of digital humanities in a completely new way, using ontological approaches.
- 4 Benardou, A., Champion, E., Dallas, C. and Hughes, L. (eds.). (2018). *Cultural heritage infrastructures in digital humanities*. Abingdon: Routledge.
This volume is an essential overview of the development and use of cultural heritage infrastructures in digital humanities, and the introductory chapter offers a perspective on the use of digital heritage for research.

References

- Anderson, S. and Blanke, T. (2012). Taking the long-view: From e-science humanities to humanities digital ecosystems. *Historical Social Research* 37(3): 147–164.
- Bass, A., Fairlee, J., Fox, K. and Sullivan, J. (2005). *The information behavior of scholars in the humanities and social sciences*. Washington, DC: University of Washington, pp. 1–14.
- Benardou, A., Constantopoulos, P. and Dallas, C. (2013). An approach to analysing working practices of research communities in the humanities. *International Journal of Humanities and Arts Computing* 7(1–2): 105–127.
- Bia, A. (2005). The digital humanities workbench. In *Proceedings of the 13th international conference on Interacción Persona-Ordenador* [online]. doi:10.1145/2379636.2379685.
- Bradley, A.J., Carpendale, S., Hancock, M. and Kirton, T. (2016). Language DNA: Visualizing a language decomposition. *Digital Humanities Quarterly* 10(4). Available at: <http://digitalhumanities.org/dhq/vol/10/4/000259/000259.html>.
- Bulger, M., et al. (2011). Reinventing research? Information practices in the humanities. *Information practices in the humanities, a research information network report*, pp. 1–84. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1859267.
- Case, D. (2012). *Looking for information: A survey of research on information seeking, needs and behaviour*, 3rd ed. Bingley: Emerald Group Publishing.
- Chumtong, J. and Kaldevey, D. (2017). *Beyond the Google Ngram viewer: Bibliographic databases and journal archives as tools for the quantitative analysis of scientific and meta-scientific concepts*. FIW Working Paper No. 8 [online]. Available at: www.fiw.uni-bonn.de/publikationen/FIWWorkingPaper/fiw-working-paper-no.-8.
- Cordell, R. (2015). Reprinting, circulation, and the network author in antebellum newspapers, in American literary history. *American Literary History* 27(3), August [online]. Available at: <http://ryan.cordell.org/research/reprinting-circulation-and-the-network-author-in-antebellum-newspapers/>.
- Council of European Social Science Data Archives – CESSDA. (2012). Available at: www.cessda.org/.
- Crane, G., Seales, B. and Terras, M. (2009). Cyberinfrastructure for classical philology. Changing the center of gravity: Transforming classical studies through cyberinfrastructure. *Digital Humanities Quarterly* 3(1): 1–38 [online]. Available at: www.digitalhumanities.org/dhq/vol/3/1/000023/000023.html.
- Education for Change Ltd, SIRU. (2002). *Researchers' use of libraries and other information sources: Current patterns and future trends* [online]. The Research Partnership, University of Brighton, pp. 1–84. Available at: www.rslg.ac.uk/research/libuse/LUrep1.pdf.
- Electronic Archiving System of Data Archiving and Networked Services – EASY. (2013). Available at: <https://easy.dans.knaw.nl/ui/home>.
- Ell, P.S. and Hughes, L.M. (2013). E-Infrastructure in the humanities. *International Journal of Humanities and Arts Computing* 7(1–2): 24–40. Available at: <http://eprints.gla.ac.uk/117395/>.
- Europeana. (2017). *Vocabulary*. Unpublished report.
- Great Britain. Department for Digital, Culture, Media and Sport. (2018). *Cultural is digital*, pp. 1–64 [online]. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/687519/TT_v4.pdf.
- Griffiths, R., Dawson, M. and Rascoff, M. (2006). Scholarly communications in the history discipline. *ITHAKA report commissioned by JSTOR*, pp. 1–14 [online]. Available at: www.sr.ithaka.org/download/file/fid/485.
- Hall, G. (2012). Has critical theory run out of time for data-driven scholarship? In M. Gold (ed.), *Debates in the digital humanities*. Minneapolis: University of Minnesota Press, pp. 128–132.
- Hitchcock, T. (2013). Confronting the digital: Or how academic history writing lost the plot. *Cultural and Social History* 10(1): 9–23.
- Houghton, J., Steele, C. and Henty, M. (2004). Research practices and scholarly communication in the digital environment. *Learned Publishing* 17: 231–232.
- Hughes, L.M. (2011). ICT methods for digital collections research. In L.M. Hughes (ed.), *Digital collections: Use, value and impact*. London: Facet.

- Hughes, L.M., Constantopoulos, P. and Dallas, C. (2015). Digital methods in the humanities: Understanding and de-scribing their use across the disciplines. In S. Schreibman, R. Siemens and J.M. Unsworth (eds.), *A new companion to digital humanities*. Oxford: Wiley-Blackwell, pp. 150–170.
- Jones, K. and Bennett, R. (2005). E-resources for research in the humanities and social sciences: A British academic policy review. *British Academy*: 1–116.
- Kirschenbaum, M. (2012). What is digital humanities and what's it doing in English departments? In M. Gold (ed.), *Debates in digital humanities*. Minneapolis: University of Minnesota, pp. 3–11.
- Liu, A. (2012). Where is cultural criticism in the digital humanities? In M. Gold (ed.), *Debates in the digital humanities*. Minneapolis: University of Minnesota Press, pp. 490–509.
- Meho, L. and Tibbo, H. (2003). Modeling the information-seeking behavior of social scientists: Ellis's study revisited. *Journal of the American Society for Information Science and Technology* 54(6): 570–587.
- Nichols, D. and Clark, D. (2015). Finding stuff. In M. Moss, B. Endicott-Popovsky and M. Dupuis (eds.), *Is digital different*. London: Facet, p. 23.
- Rayson, P.E., Mariani, J.A., Anderson-Cooper, B., Baron, A., Gullick, D.S., Moore, A. and Wattam, S. (2017). Towards interactive multidimensional visualisations for corpus linguistics. *Journal for Language Technology and Computational Linguistics* 31(1): 27–49 [online]. Available at: <https://pdfs.semanticscholar.org/0a44/c948e2afcdf51556e56a2d9c356b55f93d8e.pdf>.
- Robinson, P. (2014). Digital humanities: Is bigger, better? In P.L. Arthur and K. Bode (eds.), *Advancing digital humanities: Research methods and theories*. London: Palgrave Macmillan, pp. 243–257.
- Rosenzweig, R. (2010). *Clio wired: The future of the past in a digital age*. New York: Columbia University Press.
- Rutner, J. and Schonfeld, R. (2012). Supporting the changing research practices of historians. *ITHAKA S+R*: 1–52 [online]. Available at: www.sr.ithaka.org/sites/default/files/reports/supporting-the-changing-research-practices-of-historians.pdf.
- Service integration, data interconnect research and education – Isidore. (2013). Available at: www.rechercheisidore.fr/.
- Thaller, M. (ed.). (2012). Controversies around the Digital Humanities: An Agenda. *Historical Social Research* 37(3).
- Unsworth, J. (2000). Scholarly primitives: What methods do Humanities researchers have in common, and how might our tools reflect this? In K. College (ed.), *Symposium on humanities computing* [online]. Available at: <http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html>.
- Unsworth, J. (2003). *Tool-time, or 'haven't we been here already?* Presented at the Transforming Disciplines: The Humanities and Computer Science. Washington, DC [online]. Available at: <http://people.virginia.edu/~jmu2m/carnegie-ninch.03.html>.
- Virtual Research Environment for the Humanities – TextGrid*. (2013). Available at: www.textgrid.de/.
- Warwick, C., Terras, M., Galina, I., Huntington, P. and Pappa, N. (2008). Library and information resources and users of digital resources in the humanities. *Program: Electronic Library and Information Systems* 42(1): 5–27.
- Wulf, W.A. (1998). *Federal funding for information technology research*. Submitted to the Subcommittee on Basic Research, Committee on Science, U. S. House of Representatives. Available at: www.nae.edu/8650/FederalFundingforInformationTechnologyResearch.

English language and social media

Caroline Tagg

Introduction

Research into English language and social media (ELSM) is important to the ongoing development and conceptualisation of digital humanities, not only because it illustrates the role of digital tools in humanities research but also because it draws attention to the way in which digital technologies are transforming what it means to be human. What tends to unite researchers of language and social media is an interest in understanding the ways in which communication is increasingly shaped by the availability and use of various social media platforms from WhatsApp to YouTube. In other words, the interest is in how digital technologies shape the particular object of study, that of language (and of languages, including the English language), albeit in different ways and from various research perspectives. However, despite an increasing recognition within digital humanities research of the transformative power of technology in shaping not only how research is conducted but the outcomes of the research process itself, a distinction still tends to be made between research which uses digital technologies to study the humanities (that is, digital humanities as traditionally conceived) and that which uses humanities to study the digital. This chapter probes this distinction, showing how ELSM research (as an example of the latter) contributes to the digital humanities paradigm through the transformative impact that studying naturally occurring digital texts can have on our understanding of contemporary language use. In illustrating this argument, I focus particularly on the ways in which social media opens up access to ‘networked resources’, with implications for what it means to communicate online ‘in English’ and for who can be considered an English language user in social media contexts.

The chapter starts by giving a brief background to research into ELSM before exploring the relevance of different strands of this research for the digital humanities paradigm, focusing on the use of digital technologies as tools for collecting and visualising language data, as well as research which foregrounds how social media is transforming language as the object of study and which often leads to a rethinking of traditional, pre-digital analytical frameworks. It then moves on to focus on one particular transformation, that of the conceptualisation of English as one set of resources within individuals’ wider repertoire of multilingual and multimodal resources. A sample analysis of WhatsApp messages serves to

illustrate the ways in which English language practices online involve the production, selection and deployment of various multimedia resources and how this contributes to changing ideas about language. The implications for digital humanities in acknowledging how the humanities informs the digital are then explored.

Background

Research into ELSM falls within the well-established and rapidly growing field of language and new media or social media (e.g. Georgakopoulou and Spilioti 2015). Studies which fall into this field are defined primarily not by method, data or starting assumptions, but rather by the *medium* through which language is being used (that is, through social media). The term ‘social media’ – employed in academia to refer broadly to ‘online environments which enable social interaction’ (Leppänen et al. 2014: 113) – came into use with the proliferation of Internet platforms in the first decade of the twentieth century: social network and media sharing sites such as Facebook, Twitter, Flickr, YouTube and (more latterly) Instagram, Tumblr and Snapchat, as well as mobile messaging apps such as WhatsApp and WeChat. The studies encompassed under the umbrella term of language and social media are therefore multifarious, encompassing most, if not all, of these platforms and ranging from linguistic descriptions of particular digital language varieties (Baron 1998) to large corpus analysis of publicly available data, such as on Twitter (Zappavigna 2012) and eBay (Knight et al. 2015), and to small-scale in-depth ethnographies of people’s online experiences (Androutsopoulos 2014). A wide variety of language practices have been investigated, from punctuation (Squires 2012) to online aggression (Angouri and Tseliga 2012), and from online consumer reviews (Vasquéz 2014) to virtual business communication (Darics and Koller 2018). These studies document the many ways in which social media is transforming what it means to participate in human society, and particularly how people communicate and negotiate social relationships. Key transformations include the collapse of time and space as digital media enable far-flung interactants to communicate synchronously (Alonso and Oiarzabal 2010); the increasing ability for busy people on the move to keep in constant communication with others across work, social and personal spheres (Papacharissi 2011); and everyday users’ increasing access to an array of multimedia resources and commercial mainstream artefacts, in part because of the convergence of grassroots and corporate organisations into one space (Jenkins 2006).

Within this wider field, the focus on the English language (ELSM) has been a controversial one. In early studies of social media and language, the initial interest was often in exploring how the Internet was shaping language by identifying and describing online varieties such as the language of email (Baron 1998; Gains 1999), Internet Relay Chat (Werry 1996) or bulletin boards (Collot and Belmore 1996). The majority of these studies looked at English-language varieties, in part because the Internet and World Wide Web had been developed in English-speaking countries and did not initially support other languages or scripts (Pargman and Palme 2009), but also because of a tardiness within the academic community to look beyond English (Danet and Herring 2007: 5). Since the publication of Danet and Herring’s *The Multilingual Internet* in 2007, the implicit assumption that the language of the Internet is English has been challenged by researchers consciously charging themselves with addressing the linguistic imbalance and recognising the multilingual nature of the Internet (e.g. Lee 2016). Nonetheless, it could be argued that English continues to hold a special status as an online *lingua franca*, used alongside other languages in a range of social media contexts by speakers of various languages.

Although ELSM scholars tend not to position themselves as working within digital humanities, the work of the former is often very relevant to the varied concerns of the latter. Given the fact that work in the humanities is increasingly ‘being done digitally’ (Fitzpatrick 2012), digital humanities is rapidly expanding in scope (Klein and Gold 2016) to the point where questions are being raised as to the usefulness of a separate ‘digital humanities’ field (Deegan and Hayler 2016) or the possibility of demarcating it (Terras 2016). Digital humanities is usually defined as ‘not a unified field but an array of convergent practices’ (Schnapp et al. 2009) or a ‘conjunction and collision of many fields’ (Liu 2016: 1550) which engages in ‘ongoing, vigorous discussion of disciplinary identity’ (Liu 2016: 1547). Despite the various disciplines encompassed by digital humanities, as an umbrella term it carries with it certain assumptions and values, such as the endeavour to ‘reshape and reinvigorate contemporary arts and humanities practices’ (Schnapp et al. 2009: 13) and to explicitly reflect on the ‘difference that the digital can make to the kinds of work that we do’ (Fitzpatrick 2012: npg). At the same time, digital humanities has moved from a focus simply on ‘technology as a tool’ (Svensson 2010: npg), typically involving large-scale digitisation projects, towards an interest in interpreting ‘knowledge that is ‘born digital’ and lives in various digital contexts’ (Presner 2010: 6) alongside a growing awareness of the ways in which the digital may be transforming the object of study (Fitzpatrick 2012). Digital humanities work is transformative in the sense that scholars using digital technology to understand an artefact, practice or phenomenon find that their understanding is transformed by the perspective enabled by the technology. Collins (2016) gives the example of how the digitisation of medieval manuscripts enables present-day scholars to adopt flexible and ‘vertical’ approaches to fifteenth-century texts (that is, practices which do not rely on linear reading) that more accurately reflects the way manuscripts were produced and consumed at the time, before particular ways of interacting with texts were imposed by print (see Deegan and Hayler 2016 for other examples). Within English language studies, another example is that of corpus linguistics. From the 1960s, the use of computer technology to collect, store and interrogate large amounts of textual data has not only allowed researchers to study more texts, but the practice of looking at a large amount of data at the same time has led to new understandings of how language (initially English) works (Sinclair 1991: 100), namely in lexical patterns rather than a grammatical slot-and-filler approach. Based on corpus analysis, language production is now seen as being driven primarily by the collocational and patterning behaviour of different lexis, rather than the slotting of words into existing grammatical structures. Thus, the use of corpus technology transformed linguists’ understanding of the object of study, language itself. Key to digital humanities work, then, is the recognition that digital technology not only transforms how research is carried out but can be transformative in shaping disciplinary knowledge.

This kind of impact can be seen within strands of ELSM research for which digital technologies serve as a tool for obtaining naturally occurring language data. For example, sociolinguists who study regional patterns of linguistic variation and change are increasingly exploiting social media (and particularly Twitter) for its wealth of time-stamped, geocoded data (e.g. Bauman et al. 2014; Huang et al. 2016) in order to obtain information on language use which can otherwise only be obtained at a much smaller scale through elicitation (i.e. through interview or surveys). With a focus on lexical innovation – the emergence of new words – Grieve et al. (2018) exploit a multi-billion-word corpus of American tweets to map the spread of new forms such as *baeless* (a single person), *rekt* (wrecked, intoxicated or defeated) and *yaas* (yes) across the United States, drawing also on multivariate spatial analysis to visualise the findings. Their use of big data and digital visualisation tools enables

insights into patterns of lexical diffusion not otherwise accessible and which serve to confirm predicted patterns of diffusion – such as, for example, the role of large densely populated urban areas in lexical innovation. At the same time as confirming general language trends, however, the research also reveals how the use of social media itself shapes patterns of lexical innovation, with Atlanta in the Deep South emerging as an important origin of new forms because of the city's African American cultural influence and the relatively high engagement of African Americans on Twitter (Grieve et al. 2018: 309). The researchers conclude that, on Twitter, the cultural profile of an urban area is more important for lexical innovation and diffusion than its size or population density. It is also evident that many of the new forms on Twitter (such as *rekt* and *yaas*) are characterised by phonetic spelling and other practices primarily associated with (text-based) social media (Eisenstein 2015). In short, then, this research attempts to understand general human language and communication patterns not only through the use of language data that are conveniently produced in digital form but also by exploring the impact that digital mediation has on the language being used.

This leads to the argument underpinning this chapter, which is that research into ELSM is important not only because it shows how digital methodologies and tools can impact on the gaze of the researcher (with implications for how an object of study is conceptualised) but also because it illustrates how digital technologies can transform language use, the object of study, creating objects that are themselves 'born digital'. This line of work is not typically considered to fall within the paradigm of digital humanities. However, it could be argued that the use of humanities to understand digitally mediated interactions constitutes another facet of the complex relationships between computing and the humanities (Rosenbloom 2012). In her discussion of the boundaries of digital humanities, Fitzpatrick (2012: npg) points to the 'overlap . . . between scholars who use digital technologies in studying traditional humanities objects [i.e. "digital humanities" typically conceived] and those who use the methods of the contemporary humanities in studying digital objects'; in other words, those who use humanities to explore how digital technologies are transforming human life. Part of this overlap, as I shall argue later, can be found in the ways in which the study of digital objects goes on to transform disciplinary knowledge and research perspectives.

Critical issues and topics

A number of strands of ELSM research, both quantitative and qualitative, foreground the ways in which social media is transforming (English) language use, the object of their study. For example, drawing on the established framework of systemic functional linguistics (SFL), which emphasises the ideational and interpersonal functions of language (Halliday 1978), Zappavigna (2012, 2014) aims to illuminate a new kind of communication which she terms 'searchable talk'. Searchable talk captures the ways in which people communicating via a large fast-moving social media site like Twitter seek to create 'ambient affiliation' – to bond, however fleetingly, with other users without directly interacting with them – which they do by creating alignment through social media affordances such as hashtags (#example) and usernames (@example). In order to explore this, Zappavigna works with large Twitter corpora, created using the Twitter streaming application programming interface (API), which she explores through SFL by means of discourse analysis. A similar combination of digital tools and analytical frameworks taken from the humanities is employed by Page (2018) in her study of online stories. Page draws on and extends existing narrative analysis frameworks in situating what she calls the 'shared story' as a distinctive narrative genre facilitated

by social media platforms including Wikipedia and YouTube. Characterised by their potential to involve thousands of interactions and tellers, the shared story has offline antecedents but differs both functionally and formally from typical everyday stories told in face-to-face or private online interactions (Page 2018).

The particular ways in which communication unfolds on social media often requires researchers to revisit existing theoretical and analytical frameworks with implications for our wider understanding of language use. Jones (forthcoming, 2020), for example, explores the implications of online algorithms for traditional pragmatics, arguing that the ways in which users interact with algorithms – which monitor users' actions in order to create inferences about the users' future actions – requires researchers to rethink some of the fundamental principles of meaning and interaction. Similarly, Tagg et al. (2017) argue that social media requires a shift in how we conceptualise sociolinguistics frameworks of audience design and addressivity, with Facebook users needing to write themselves and their audiences into being through their posts (to paraphrase boyd 2008). Given the complexities of the online space – such as the diversity of online connections within one's Facebook friends and the fact that the particular audience for any one status update is unknown – users must more carefully style their utterances than in typical face-to-face situations, in order to target particular users, exclude others and avoid creating offence (Tagg et al. 2017). In studies positing shifts in established thinking, the argument is not always that social media creates new or completely unique forms of communication (Facebook can, in terms of its diverse audience, be compared to an offline situation like a wedding), but rather that the novelty and salience of social media draw attention to practices which have a bearing on communication more generally, potentially transforming existing ideas about language use. As Page (2018: 23) argues, for example, a focus on the non-linear and open-ended nature of online shared stories brings 'fresh impetus' to wider thinking around narrative structure.

As with much digital humanities work, then, ELSM is transforming how we understand what it means to be human, not only through the application of digital technologies but through its focus on how digital technologies are shaping the digital object, in this case the language of social media. One of the ways social media transforms languages use, I will argue later, is by making available a particular set of multimodal and networked resources. As we shall see, the act of analysing how English is used through social media has not only informed ELSM researchers as to how the English language is being used in the twenty-first century but also feeds into transformative ideas regarding the English language, and language more generally.

Current contributions and research

Social media has radically altered the possibilities for human communication and the ways in which language, including the English language, is used. One of the key transformations has been the opening up of the resources available to users and the increasing potential for everyday users to contribute to cultural artefacts on sites like YouTube through, for example, remixing and sampling (Leppänen and Elo 2016), and to public conversations on Twitter through, for example, the use of hashtags (Zappavigna 2012) and retweets (Draucker and Collister 2015). The availability of multimedia (through access to different platforms and apps) and multimodal resources – that is, an array of different modes including image, video, emoji and GIFs – is among the more salient aspects of social media communication, as is the increased potential for users to move between and combine different languages.

As mentioned earlier, the research community was slow to recognise the multilingual potential of the Internet, with Danet and Herring (2007) at the forefront of studies of languages other than English. More recent studies of online language use find ‘unprecedented’ possibilities for language contact online (Androutsopoulos 2013: 667), often involving English, due to the way in which the Internet facilitates communities which span geographical distance and cultural divides. Scholars note a great deal of innovative code switching and script mixing across digital mediums as well as ‘trans-scripting’ (Androutsopoulos 2015) – writing a language using the spelling conventions of another, as when French *bisou* (‘kiss’) is rewritten according to English spelling conventions as <bizoo> (Morel et al. 2014: 274). The creative Romanisation of languages such as Chinese (Su 2007), Arabic (Palfreyman and Al Khalil 2007) and Greek (Spilioti 2009) also shows how a global language such as English can influence – and be influenced by – other languages on social media. This research has thus contributed in important ways to emerging ideas in sociolinguistics regarding the fluid, often unpredictable boundaries between languages. These ideas, predominantly associated with spoken face-to-face communication, are grounded in the empirical observation that language users do not, in practice, speak one language or another, but instead draw more fluidly on a range of named languages, language varieties, registers and styles in their individual repertoires – a practice known by some researchers as ‘translanguaging’ (e.g. Creese and Blackledge 2010; Garcia and Wei 2014). Social media has been recognised as playing a central role in facilitating such practices by connecting far-flung individuals and communities (Blommaert and Rampton 2011). Research into online multilingualism adds to this body of research by showing the implications for translanguaging when people have access to ‘a flow or semiotic [or communicative] resources that by far exceeds the resources that individual members bring along’ (Androutsopoulos and Juffermans 2014: 4); in other words, when people have access to additional resources available through their networks.

The expansion of resources available to Internet users is a significant technological shift. There was an initial tendency to assume – among researchers as well as Internet users and the media – that computer-mediated communication was dry and impoverished (Thurlow et al. 2004: 45–57). This was theoretically conceptualised with reference to ‘reduced social cues’ (Sproull and Kiesler 1986), the argument being that the lack of ‘contextualisation cues’ (Gumperz 1982) in the largely text-based online communication of the late twentieth century decreased inhibitions and led to online aggression; while, in the mainstream media, perceptions around the impoverished nature of online writing led to claims regarding the detrimental effect that technology was having on language: ‘the English language as we once knew it is out the window’ (Uthus 2007) declared one newspaper. This assumption overlooked the creative and personal ways in which interlocutors sought to express themselves online with, for example, the emoticon or smiley being invented in 1982 by Scott E. Fahlman at Carnegie Mellon University in order for forum users to disambiguate meaning and indicate humour and irony (Baron 2009). The inherently playful nature of the Internet – encouraged rather than discouraged by affordances such as anonymity, lack of physical presence and other communicative constraints – was quickly recognised in a series of studies looking at humour and play online (Bechar-Israeli 1995; Danet et al. 1997; Reid 1991), and this line of research continues today (Deumert 2014; Thurlow 2012).

This shift in perspective regarding the creative potential of online communication has been accompanied by developments in technology which mean that the resources available to early Internet users – written words, respelling and typographic resources such as punctuation (the constraints of which in fact encouraged typographical play) – have now greatly expanded. Social media users today have access to an array of material and resources which

were not similarly available to many people in the pre-Internet era – these include photos and other images, video, audio, GIF and emoji. Social media communications are thus increasingly multimodal, by which meaning is produced through drawing on and combining a range of different semiotic modes including font, image and colour (Kress and van Leeuwen 1996). Social media users have access not only to the graphic resources afforded by the computer or mobile phone keyboard, but are also increasingly able to draw on ‘networked resources’ – that is, a range of external resources which can be linked to from ‘in the network’ and/or embedded in new sites (Androutsopoulos 2015). These networked resources can include language resources (such as dictionaries or online translations) as well as global commercial artefacts such as music videos which are now available for users to interact with and exploit (Leppänen and Elo 2016). Also relevant here is the extent to which social media are now seen as being embedded in people’s lives in a way that earlier platforms were not, with growing significance for what it means to be human. Howard Rheingold, an early Internet user, wrote of using his computer to escape his daily life and to meet a virtual community largely separate from his physical, everyday life (Rheingold 1993). Online groups continue to provide the opportunity for people to meet like-minded others with whom they would otherwise not come into contact. However, online activities are now much more entangled in people’s everyday offline activities, and people move between physical and virtual contexts for various reasons and in a variety of ways (Cohen 2015; Monaghan 2014). Social media users can take photos or videos with a mobile phone and then embed them in an online interaction or website (many mobile apps allow users to take a photo from within the app and then immediately send it), they can link to a website or Facebook page from a tweet or they can record themselves and send an audio message through WhatsApp. (Examples of these are explored in the sample analysis in the next section.) Whereas once the text-based environment of an Internet forum or online game enabled people to create new identities (Reid 1991), the contemporary online sharing of multimodal resources such as selfies or holiday photos mean that people’s online identities are often an extension of their offline ones.

Social media users can also use digital technology and editing suites to creatively design content by drawing on and reworking existing content; these result in remixes and mashups (Williams 2009), as well as buffalaxed videos – music videos on to which people add subtitling that does not match the original verbal content (Leppänen and Elo 2016). The creation of these artefacts often depends on users having access to globally circulating commercial content as well as to the tools needed to digitally remaster them. They can result in what Androutsopoulos (2010) calls ‘vernacular spectacles’ – localised performances on media-sharing sites like YouTube to which other users can contribute through comments and further remixing. As well as enabling users to engage politically and critically with mainstream content, this access to networked resources can also encourage them to create new identities and align with particular communities, as illustrated by Jonsson and Muhonen’s (2014) study of young Swedish girls whose Facebook profiles highlighted their affiliation with Japanese manga. Users can post simultaneously on different sites (the same photos onto Facebook and Instagram, for example) and create a digital biography that extends across media platforms, as Adami (2014) shows in her analysis of one blogger’s transmedia practices. As this suggests, the availability of networked resources – as well as the multiple modes afforded by social media – has important implications for our understanding of ELSM.

As mentioned earlier, access to networked resources also increases the opportunity for users to engage in multilingual practices. First, the network makes available to users affordances such as online dictionaries and translation tools and thus gives them recourse to

resources from across a greater range of languages than might be the case in face-to-face spoken interactions. Second, stickers and emoji can be used as part of a conversation without the user themselves having to produce language (Tagg 2015). Third, the fact that users can repost others' contributions, embed videos in their posts or link to other sites enables them to construct a multilingual assemblage of a kind not usually associated with spoken interaction. These practices are likely to contribute to a kind of 'minimal bilingualism' (Androutsopoulos 2007) which involves the use of accessible, formulaic phrases from a language (in online contexts, often English) which is not spoken to a level of proficiency by the individual or community involved. In their study of Swiss Short Message Service (SMS) text messaging, for example, Morel et al. (2014: 283) found that most code switching into English involved users inserting formulaic greetings and expressions such as 'you're welcome!' and 'What's up?'. Thus, what is often encouraged by users' access to networked resources are translanguaging practices which involve users drawing fluidly on resources from a range of languages in which they may not always be proficient.

The availability of networked resources therefore has implications for what it means to be an online user of English (or any other language). The use of networked resources in online conversation can be theorised using Goffman's (1981) ideas about production format. According to Goffman, the label 'speaker', as used in analysis of spoken interaction, can refer to the *animator* (the person who actually makes the utterance); the *author* (the person who composed the words uttered); and/or the *principal* (the person whose beliefs are represented by the utterance). These three roles may all be carried out by one person, but not necessarily so. For example, a spokesperson can be seen as an animator reading the words of the script writer (the author), while the politician is the principal. Applying Goffman's framework to Twitter, Draucker and Collister (2015) suggest that it can be useful to distinguish between the animator and the 'broadcaster' to describe instances in which a tweet is retweeted (or 'broadcast') by another user. In this case, the original author usually remains the animator, author and principal, although it is possible for this to be lost or obscured in a chain of citations (i.e. where a retweeted post is retweeted). Similarly, somebody who shares multimodal content might be seen as broadcasting the work of the original author. Both retweeting and sharing, like reported speech, produce what Bakhtin (1981) calls 'multi-voiced' discourse, as broadcasters inevitably take a stance on the repeated material, parodying it or (more often with retweets, according to boyd et al. 2010) aligning with the views expressed. Whatever their stance, users' access to networked resources opens up the means available for them to join a conversation and take a turn.

In the sample analysis that follows, I explore English language interactions between friends via a messaging app which illustrate some of the ways in which digital technologies are transforming our object of study, that of contemporary language use. The argument that I aim to address is whether the use of the humanities to explore and understand the digital can usefully be seen as part of the digital humanities.

Sample analysis

This section looks at the communication taking place between a network of friends via the mobile messaging app, WhatsApp. The app provides a communicative space within which dyads and groups can share short typed messages, images, videos, audio recordings and links. What emerges from this space often resembles a conversation, characterised by participant turns, initiations, question-and-answer sequences, response tokens and other features associated with spoken conversations. The conversational nature of these exchanges

is also foregrounded by the affordances of WhatsApp, including the speech bubble format and the fact that users are informed as to whether their interlocutors are online, have read their posts and are replying. At the same time, unlike face-to-face conversations, WhatsApp interactions can take place between interlocutors in different physical places and in their own time (it is, at best, ‘quasi-synchronous’, Garcia and Jacobs 1999). Furthermore, the networked resources and multimedia available through WhatsApp transform the communication and, by extension, our ideas regarding the boundaries of English as used in social media interactions.

The data explored here are taken from a large AHRC-funded ethnographic project, ‘Translation and Translanguaging: Investigating Linguistic and Cultural Transformations Across Superdiverse Wards in Four UK Cities’.¹ The project methodology involved our working closely with key participants. As well as observing and recording each key participant’s interactions at work, we asked them to make recordings at home and to contribute social media data. The social media data were in most cases collected from participants’ mobile phones through screengrabs, taken either by the participants themselves or by the research fellows. Where necessary, these were translated but we otherwise worked directly with the screengrabs, which gave us access to the original context in which the online posts were read. As will already be evident, the methodology differs in key ways from the other projects discussed earlier in this chapter. The social media data were collected only as part of a wider online ethnography, rather than constituting the main focus of the project, with the implication that the individual’s online practices could be contextualised within their wider communicative practices and relationships. The focus, then, was not primarily on how digital technology shaped language use, but rather on how particular individuals interpreted and incorporated the technology into their communicative practices. Of particular relevance is the fact that, beyond the use of screengrabs and digital anonymisation tools, the project did not use digital tools to analyse or visualise the data, and so it was not a digital humanities project as typically conceived. However, while the digital did not shape our understanding of the humanities, we used humanities – frameworks like Goffman’s as part of a linguistic ethnography – to understand the digital.

As part of this project, we worked with J., a volleyball coach and beauty salon manager in Birmingham who was originally from Hong Kong. J. submitted approximately 150 screengrabs of his mobile phone, containing around 650 WhatsApp and SMS messages. These included individual interactions with colleagues, friends and members of his volleyball team, as well as group chats with friends and with his family and friends back home. What is interesting about his interactions includes the ways in which J. and his interlocutors exploit networked resources, appropriating them as turns in their conversations. For example, as illustrated in Figure 30.1, conversations in J.’s WhatsApp groups can be initiated by website links. In Figure 30.1, J. is not the author of the utterance nor the animator as such, but rather the broadcaster – or ‘sharer’, given the private, bounded context, while the recontextualised weblink appears to serve a clear communicative function for J. in a way that positions him as principal of his turn. It is evident from G.’s response (‘so funny’) that she found the shared resource both amusing and relevant, making comparisons with her own parenting of her son, F.

Below I explore what existing analytical frameworks can reveal about this new form of communication by focusing on one longer extract taken from an interaction between J. and S., a close friend and also senior manager in J.’s salon (Figure 30.2).

Although ostensibly an English language exchange, the boundaries of what this constitutes is expanded in three main ways: through creative respellings, images and emoji.

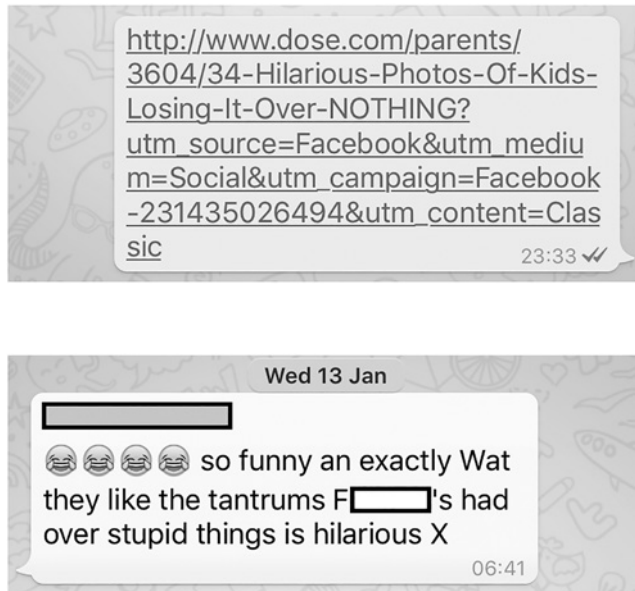


Figure 30.1 Posted in group labelled 'Eating N Drinking', four participants

I will address these three sets of resources in turn. First, the creative or unconventional spelling variations in Figure 30.2 include letter repetitions or 'flooding' in 'Awww' and 'soooooo', and <ur> (in 'How's ur beef stew?') which can be analysed either as a clipping of <your> or as a use of homophones (<u> for 'you' + <r> for 'are'). Drawing on sociolinguistic ideas about spelling as a meaningful social practice (Sebba 2007), we can see that the letter flooding emphasises the emotive evaluation which S. wishes to convey. Although <ur> is ostensibly a shortening carried out for reasons of efficiency, it is worth noting that J. makes the effort of adding an apostrophe to 'How's' (i.e. that he is not only concerned with efficiency) and therefore speculating whether <ur> is also carrying out identity or relational work (Tagg 2012). Thus, as in other new media practices, the communicative possibility of English is expanded by the potential for orthographic choices to convey meaning within an everyday online interaction (Shortis 2016).

Second, the use of images can be explored using Goffman's participation framework. J. initiates an exchange within Figure 30.2 by posting a screengrab displaying an online voucher for a 'Live Escape Game for Two to Six People at Clue HQ Birmingham (57% Off)' (a game in which people must solve clues in order to obtain a door code and escape from a locked room). Thus, the referential content of his post is conveyed not directly by J., but through the text (and image) on this screengrab, meaning that he is 'sharer' but not author or animator of his turn (Goffman 1981). By sharing the screengrab in this context, J. does not take an evaluative stance towards it, but uses it to perform a particular speech act, namely an offer or suggestion along the lines of 'Do you fancy doing this Escape Game?'. This seems to be the interpretation made by S., who responds accordingly. We might then suggest that J. is the principal of his turn, 'the party to whose position the words attest' (Goffman 1981: 226).

When J. changes the subject to that of food, he responds to S.'s question 'what you got?' with a picture of his half-finished dinner and the caption 'Groupon' (an online company that



Figure 30.2 Individual interaction with S

supplies vouchers and discounts). The photo constitutes what Zappavigna and Zhao (2017) call an ‘inferred selfie’, the nature and placement of the objects suggesting the presence of the photographer. If we thus assume that he took the photo and posted it to WhatsApp, we can argue that J. is author, animator and principal of this turn, but that each role differs from that envisaged for face-to-face interaction (the author being the photographer rather than the composer of a written utterance, for example) and that meaning is created primarily by image, with the caption complementing the image by extending the information provided (Unsworth 2004) – perhaps explaining why he is eating out. It is a common practice in J.’s WhatsApp interactions for a photo of what the poster is eating to serve as a turn, apparently fulfilling



Figure 30.2 (Continued)

a communicative function as part of an online interaction without the need for words. This form of communication is enabled by the convergence of various digital functionalities – camera, messaging app, Internet connection – into the everyday user’s mobile phone.

Third, as is evident in Figure 30.2, emoji are frequently used in these WhatsApp interactions and they serve a number of pragmatic functions (Danesi 2016). In Figure 30.2, S. uses a heart emoji to add an evaluative stance to her verbal response (‘Cool, it was amazing!’). In a subsequent turn, three heart emoji serve as a turn in their own right, where they convey an evaluative stance and also mark S.’s presence in the conversation – that is, in a medium where she cannot rely on gestures or facial expressions, she is showing that she is listening.

In this way, emoji standing on their own act as backchannels or response tokens – minimal contributions realised in speech by words like ‘yeah’ or ‘okay’. As whole turns, emoji are always multiple in this dataset, with up to six or so emoji (usually identical, but sometimes in combination) appearing alongside each other.

These pragmatic functions are further illustrated by the WhatsApp group chat in Figure 30.3. The first turn is in response to a photo posted in the group space. In the second turn, G. replies with a response token of three emoji representing laughter, before moving on to the main part of her utterance: ‘wen he gets back I will x’. After two audio recordings posted



Figure 30.3 Emoji posted in group ‘Eating N Drinking’, four participants

by D. (another illustration of the multimodal nature of the interactions in this group), S. adds two different emoji to her next turn, a pointing finger and a dancing woman, in a way that serves to illustrate and bring to life her claim that ‘We can rock the boogie’. G. then replies with a series of seven emoji that stand for her full turn. There is a sense that these interlocutors are not engaged in transactional or informational but highly interpersonal discourse and that the turns serve mainly to keep open a channel of communication (Ito and Okabe 2005) and to show users’ participation and engagement in the only way possible in social media – by making a contribution to the discourse.

These examples serve to illustrate the wide range of digital, multimodal and networked resources available to, and exploited by, interlocutors on WhatsApp; the various pragmatic and communicative functions that these resources can fulfil; and the implications for the ways in which people can participate in social media interactions, for example, as ‘sharers’ rather than animators in the sense of physically producing the text and not as authors. Although the interlocutors in the examples here are all fluent English users, this in turn has implications for our understanding of the skills and abilities which equip people to communicate in English – the global *lingua franca* – in the contemporary world. An English language user on social media may not necessarily be what is traditionally considered ‘fluent’ in English as long as they have access to the networked resources and media that enable them to animate English language comments; in other words, as long as they are aware of how to exploit the resources available to them through the network. Rather than ability in any one particular language, an individual’s communicative flexibility, language/media awareness and digital literacy skills may instead prove more important in defining a successful communicator online.

Meanwhile, the ways in which these various modes work alongside English within an online conversation, often in place of written text, serve to transform our understanding of the boundaries between language and other modes. English language utterances do not necessarily have to constitute a sender’s own words, but instead the words of others can be linked to through screengrabs and weblinks (such as the Escape Game voucher), while turns in an English language exchange can be conveyed through other modes entirely (such as image and emoji). In this way, the digital serves to transform the object of study, with implications for how existing frameworks of analysis must be altered to accommodate new practices.

Future directions

In this chapter, I focused on the impact that recognition of the availability of networked resources has had on our understanding of what it means to communicate online ‘in English’ and the implications for who can be considered an English language user in social media. For many scholars, such analyses and insights may fall outside the purview of digital humanities, at least as traditionally defined. The interest of digital humanities scholars typically lies in the transformative impact that digitalisation and digital technologies can have on the research process and, as such, are much more easily aligned with ELSM projects such as those described earlier in the chapter which use social media as a source of data and which exploit digital technologies in visualising and analysing it (e.g. Grieve et al 2018). However, what unites digital humanities research is increasingly a recognition of the transformative power of technology in shaping not only how research is conducted, but the outcomes of the research process itself. In this sense, ELSM research contributes to the digital humanities paradigm by highlighting the transformative impact that social media has on the research object and, by extension, how this informs and shapes wider scholarship.

Amidst rapid technological developments and shifts in the roles that digital technologies play, digital humanities is as much about exploring and pushing at the limits of what digital technologies can do and the role they play both within and beyond academia, than it is about carving out and investigating its own space. The relationships between the digital and the humanities are complex and ever-shifting (Rosenbloom 2012). As an ever-widening range of social, personal and work-related activities are increasingly mediated by digital technology, the need for theories and frameworks of analysis within the humanities to account for its impact on their object of study will only grow. As discussions around the identity of digital humanities continue, it may be necessary for scholars to consider how their objects of study are being transformed not only by digitalised research methods but also by the key roles that digital technologies increasingly play in human life.

Note

- 1 AHRC *Translation and Translanguaging: Investigating Linguistic and Cultural Transformations in Superdiverse Wards in Four UK Cities*. (AH/L007096/1). Angela Creese (PI). With CIs Mike Baynham, Adrian Blackledge, Frances Rock, Lisa Goodson, Li Wei, James Simpson, Caroline Tagg, Zhu Hua.

Further reading

- 1 Androutsopoulos, J. (2015). Networked multilingualism: Some language practices on Facebook and their implications. *International Journal of Bilingualism* 19(2): 185–205.
This article discusses the resources available to (multilingual) users of social media.
- 2 Androutsopoulos, J. and Staehr, A. (2018). Moving methods online: Researching digital language practices. In A. Creese and A. Blackledge (eds.), *The Routledge handbook of language and superdiversity*. Abingdon: Routledge, pp. 118–132.
This chapter discusses some of the characteristics of online text and explores the implications for research methods.
- 3 Deumert, A. (2014). *The sociolinguistics of mobile communication*. Edinburgh: Edinburgh University Press.
This book discusses creativity, intertextuality and multimodality within mobile communications from a global perspective.
- 4 Grieve, J., Nini, A. and Guo, D. (2018). Mapping lexical innovation on American social media. *Journal of English Linguistics* 46(4): 293–319.
This article, as discussed in the chapter, draws on naturally occurring digital texts to explore language variation and innovation.
- 5 Leppänen, S. and Elo, A. (2016) Buffalaxing the other: Superdiversity in action on YouTube. In K. Arnaut, J. Blommaert, B. Rampton and M. Spotti (eds.), *Language and superdiversity*. Abingdon: Routledge, pp. 110–136.
This chapter illustrates the many ways in which social media users can draw on and exploit networked resources for political and interpersonal ends.

References

- Adami, E. (2014). Retweeting, reposting, repinning; reshaping identities online: Towards a social semiotic multimodal analysis of digital remediation. *Lingue e Letteratura d'Oriente e d'Occidente* 3: 223–243.
- Alonso, A. and Oiarzabal, P.J. (eds.). (2010). *Diasporas in the new media age: Identity, politics and community*. Reno: University of Nevada Press.

- Androutsopoulos, J. (2007). Bilingualism in the mass media and on the internet. In M. Heller (ed.), *Bilingualism: A social approach*. New York: Palgrave Macmillan, pp. 207–230.
- Androutsopoulos, J. (2010). Localising the global on the participatory web. In N. Coupland (ed.), *The handbook of language and globalisation*. Oxford: Wiley-Blackwell, pp. 203–231.
- Androutsopoulos, J. (2013). Code-switching in computer-mediated communication. In S.C. Herring, D. Stein and T. Virtanen (eds.), *Pragmatics of computer-mediated communication*. Berlin and New York: Mouton de Gruyter, pp. 659–686.
- Androutsopoulos, J. (2014). Moments of sharing: Entextualisation and linguistic repertoires in social networking. *Journal of Pragmatics* 73: 4–18.
- Androutsopoulos, J. (2015). Networked multilingualism: Some language practices on Facebook and their implications. *International Journal of Bilingualism* 19(2): 185–205.
- Androutsopoulos, J. and Juffermans, K. (2014). Digital language practices in superdiversity: Introduction. *Discourse, Context and Media* 4–5: 1–6.
- Angouri, J. and Tseliga, T. (2012). ‘You have no idea what you are talking about!’ From e-disagreement to e-impoliteness in two online fora. *Journal of Politeness Research* 6: 57–82.
- Bakhtin, M. (1981). *The dialogic imagination* (C. Emerson and M. Holquist, trans.). Austin, TX: University of Texas Press.
- Baron, N. (1998). Letters by phone or speech by other means: The linguistics of email. *Language and Communication* 18: 133–170.
- Baron, N. (2009). The myth of impoverished signal: Dispelling the spoken-language fallacy for emoticons in online communication. In J. Vincent and L. Fortunati (eds.), *Electronic emotion: The mediation of emotion via information and communication technologies*. London: Peter Lang, pp. 107–136.
- Bauman, D., Eisenstein, J. and Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18: 135–160.
- Bechar-Israeli, H. (1995). From bonehead to clonehead: Nicknames, play, and identity on Internet relay chat. *Journal of Computer-Mediated Communication* 1(2).
- Blommaert, J. and Rampton, B. (2011). *Language and superdiversity: A position paper*. Working Papers in Urban Language & Literacies (Paper 70).
- boyd, d. (2008). *Taken out of context: American teen sociality in networked publics*. PhD dissertation, University of California-Berkeley, School of Information.
- boyd, d., Golder, S. and Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In *Proceedings of the 43rd Hawaii international conference on system sciences 2010*. Los Alamitos, CA: IEEE Press, pp. 1–10.
- Cohen, L. (2015). World attending in interaction: Multitasking, spatializing, narrativizing with mobile devices and Tinder. *Discourse, Context and Media* 9: 46–54.
- Collins, M. (2016). *Caxton’s apprentice: Editorial intervention and digitising historical texts*. Paper presented at Historicising the Digital, University of Leicester, 27–28 June.
- Collot, M. and Belmore, N. (1996). Electronic language: A new variety of English. In S.C. Herring (ed.), *Computer-mediated communication*. Amsterdam: John Benjamins, pp. 13–28.
- Creese, A. and Blackledge, A. (2010). Translanguaging in the bilingual classroom: A pedagogy for learning and teaching? *The Modern Language Journal* 94(i): 103–115.
- Danesi, M. (2016). *The semiotics of emoji: The rise of visual language in the age of the internet*. London: Bloomsbury.
- Danet, B. and Herring, S.C. (2007). *The multilingual internet: Language, culture, and communication online*. Oxford: Oxford University Press.
- Danet, B., Ruedenberg-Wright, L. and Rosenbaum-Tamari, Y. (1997). ‘Hmmm . . . where’s that smoke coming from?’ Writing, play and performance on Internet relay chat. *Journal of Computer-Mediated Communication* 2(4).
- Darics, E. and Koller, V. (2018). *Language in business, language at work*. London: Palgrave Macmillan.
- Deegan, M. and Hayler, M. (2016). Digital humanities and the future of the book. In A. Hewings, L. Prescott and P. Seargeant (eds.), *Futures for English studies: Teaching language, literature and creative writing in higher education*. London: Palgrave Macmillan, pp. 161–178.

- Deumert, A. (2014). The performance of a ludic self on social network(ing) sites. In P. Seargeant and C. Tagg (eds.), *The language of social media: Identity and community on the internet*. Basingstoke: Palgrave Macmillan, pp. 23–45.
- Drauker, F. and Collister, L. (2015). Managing participation through modal affordances on Twitter. *Open Library of Humanities* 1(1): 1–36.
- Eisenstein, J. (2015). Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics* 19(2): 161–188.
- Fitzpatrick, K. (2012). *The humanities, done digitally*. Debates in the Digital Humanities. Available at: <http://dhdebates.gc.cuny.edu/debates/text/30> (Accessed 25 October 2016).
- Gains, J. (1999). Electronic mail – a new style of communication or just a new medium?: An investigation into the text features of e-mail. *English for Specific Purposes* 18: 81–100.
- Garcia, A.C. and Jacobs, J.B. (1999). The eyes of the beholder: Understanding the turn-taking system in quasi-synchronous computer-mediated communication. *Research on Language and Social Interaction* 32(4): 337–367.
- Garcia, O. and Wei, L. (2014). *Translanguaging: Language, bilingualism and education*. London: Palgrave.
- Georgakopoulou, A. and Spilioti, T. (eds.). (2015). *The Routledge handbook of language and digital communication*. Abingdon: Routledge.
- Goffman, E. (1981). *Forms of talk*. Philadelphia: University of Pennsylvania Press.
- Grieve, J., Nini, A. and Guo, D. (2018). Mapping lexical innovation on American social media. *Journal of English Linguistics* 46(4): 293–319.
- Gumperz, J. (1982). *Discourse strategies*. Cambridge: Cambridge University Press.
- Halliday, M.A.K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. London: Edward Arnold.
- Huang, Y., Guo, D., Kasakoff, A. and Grieve, J. (2016). Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59: 244–255.
- Ito, M. and Okabe, D. (2005). Technosocial situations: Emergent structurings of mobile email use. In M. Ito, D. Okabe and M. Matsuda (eds.), *Personal, portable, pedestrian: Mobile phones in Japanese life*. Cambridge: MIT Press, pp. 257–276.
- Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. New York: New York University Press.
- Jones, R. (forthcoming 2020). The rise of the pragmatic web: Implications for rethinking meaning and interaction. In C. Tagg and M. Evans (eds.), *Message and medium: English language practices across old and new media*. Berlin: De Gruyter Mouton.
- Jonsson, C. and Muhonen, A. (2014). Multilingual repertoires and the relocalization of manga in digital media. *Discourse, Context and Media* 4–5: 87–100.
- Klein, L.F. and Gold, M.K. (2016). Digital humanities: The expanded field. In *Debates in the digital humanities*. Minneapolis: University of Minnesota Press.
- Knight, D., Walsh, S. and Papagiannidis, S. (2015). I’m having a spring clear out: A corpus-based analysis of e-transactional discourse. *Applied Linguistics* 38(2): 234–257.
- Kress, G. and van Leeuwen, T. (1996). *Reading images: The grammar of visual design*. London: Psychology Press.
- Lee, C. (2016). *Multilingualism online*. Abingdon: Routledge.
- Leppänen, S. and Elo, A. (2016). Buffalaxing the other: Superdiversity in action on YouTube. In K. Arnaut, J. Blommaert, B. Rampton and M. Spotti (eds.), *Language and superdiversity*. Abingdon: Routledge, pp. 110–136.
- Leppänen, S., Kytölä, S., Jousmaki, H., Peuronen, S. and Westinen, E. (2014). Entextualisation and resemiotization as resources for identification in social media. In P. Seargeant and C. Tagg (eds.), *The language of social media: Identity and community online*. Basingstoke: Palgrave Macmillan, pp. 112–136.
- Liu, A. (2016). Is digital humanities a field? – An answer from the point of view of language. *Journal of Siberian Federal University* 7: 1546–1552.

- Monaghan, F. (2014). Seeing red: Social media and football fan activism. In P. Seargeant and C. Tagg (eds.), *The language of social media: Identity and community on the internet*. London: Palgrave, pp. 228–254.
- Morel, E., C. Bucher, S. Pekarek Doehler and B. Siebenhaar (2014). SMS communication as plurilingual communication: Hybrid language use as a challenge for classical code-switching categories. In L.-A. Cougnon and C. Fairon (eds.), *SMS communication: A linguistic approach*. Amsterdam: John Benjamins, pp. 111–140.
- Page, R. (2018). *Narratives online: Shared stories in social media*. Cambridge: Cambridge University Press.
- Palfreyman, D. and Al Khalil, M. (2007). “A funky language for teenzz to use”: Representing Gulf Arabic in instant messaging. In B. Danet and S.C. Herring (eds.), *The multilingual internet: Language, culture, and communication online*. Oxford: Oxford University Press, pp. 43–63.
- Papacharissi, Z. (2011). *A networked self: Identity, community, and culture on social network sites*. New York: Routledge.
- Pargman, D. and Palme, J. (2009). ASCII imperialism. In M. Lampland and S.L. Star (eds.), *Standards and their stories: How quantifying, classifying and formalizing practices shape everyday life*. New York: Cornell University Press, pp. 177–199.
- Presner, T. (2010). *Digital humanities 2.0: A report on knowledge*. Available at: <http://cnx.org/content/m34246/1.6/?format=pdf> (Accessed 31 May 2018).
- Reid, E.M. (1991). *Electropolis: Communication and community on internet relay chat*. PhD thesis, University of Melbourne.
- Rheingold, H. (1993). *The virtual community: Understanding on the electronic frontier*. Available at: www.rheingold.com/vc/book/ (Accessed 14 January 2014).
- Rosenbloom, P.S. (2012). Towards a conceptual framework for the digital humanities. *Digital Humanities Quarterly* 6(2).
- Schnapp, J., Presner, T., Lunenfeld, P., et al. (2009). *The digital humanities manifesto 2.0. A digital humanities manifesto, UCLA digital humanities*, 29 May. Available at: <http://manifesto.humanities.ucla.edu/2009/05/29/the-digital-humanities-manifesto-20/> (Accessed 25 October 2016).
- Sebba, M. (2007). *Spelling and society: The culture and politics of orthography around the world*. Cambridge: Cambridge University Press.
- Shortis, T. (2016). Texting and other messaging: Written system in digitally mediated vernaculars. In V. Cook and D. Ryan (eds.), *The Routledge handbook of the English writing system*. Abingdon: Routledge, pp. 487–511.
- Sinclair, J. (1991). *Corpus concordance collocation*. Oxford: Oxford University Press.
- Spilioti, T. (2009). Graphemic representation of text-messaging: Alphabet choice and code-switches in Greek SMS. *Pragmatics* 19(3): 393–412.
- Sproull, L. and Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communication. *Management Science* 32: 1492–1512.
- Squires, L. (2012). Who’s punctuating what? Sociolinguistic variation in instant messaging. In Jaffe, A.J. Androutsopoulos, M. Sebba and S. Johnson (eds.), *Orthography as social action: Scripts, spelling, identity and power*. Berlin: De Gruyter Mouton, pp. 289–324.
- Su, H.-Y. (2007). The multilingual and multiorthographic Taiwan-based internet: Creative uses of writing systems on college-affiliated BBSs. In B. Danet and S.C. Herring (eds.), *The multilingual internet: Language, culture, and communication online*. Oxford: Oxford University Press, pp. 64–86.
- Svensson, P. (2010). The landscape of digital humanities. *Digital Humanities Quarterly* 4(1). Available at: <http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html> (Accessed 25 October 2016).
- Tagg, C. (2012). *The discourse of text messaging*. London: Continuum.
- Tagg, C. (2015). *Language, business and superdiversity: A report on social media across case studies*. Working Papers in Translanguaging and Translation (WP 6).
- Tagg, C., Seargeant, P. and Brown, A. (2017). *Taking offence on social media: Communication and conviviality on Facebook*. London: Palgrave.

- Terras, M. (2016). A decade in digital humanities. *Journal of Siberian Federal University* 7: 1637–1650.
- Thurlow, C. (2012). Determined creativity: Language play in new media discourse. In R. Jones (ed.), *Discourse and creativity*. London: Pearson, pp. 169–190.
- Thurlow, C., Lengel, L. and Tomic, A. (2004). *Computer mediated communication: Social interaction and the internet*. London: Sage.
- Unsworth, L. (2004). Comparing school science explanations in books and consumer-based formats: The role of images, image/text relations and hyperlinks. *International Journal of Instructional Media* 31(3): 283–301.
- Uthus. (2007). Text messages destroying our language. *The Daily of the University of Washington*, 7 May.
- Vásquez, C. (2014). ‘Usually not one to complain but . . .’: Constructing identities in user-generated online reviews. In P. Seargeant and C. Tagg (eds.), *The language of social media: Identity and community online*. Basingstoke: Palgrave Macmillan, pp. 65–90.
- Werry, C. (1996). Linguistic and interactional features of internet relay chat. In S.C. Herring (ed.), *Computer-mediated communication*. Amsterdam: John Benjamins, pp. 47–64.
- Williams, B.T. (2009). *Shimmering literacies: Popular culture and reading and writing online*. Bern: Peter Lang.
- Zappavigna, M. (2012). *Discourse of Twitter and social media: How we use language to create affiliation on the web*. London: Bloomsbury.
- Zappavigna, M. (2014). CoffeeTweets: Bonding around the bean on Twitter. In P. Seargeant and C. Tagg (eds.), *The language of social media: Identity and community on the internet*. London: Palgrave Macmillan, pp. 139–160.
- Zappavigna, M. and Zhao, S. (2017). Selfies in ‘mommyblogging’: An emerging visual genre. *Discourse, Context & Media* 20: 239–247.

Index

Note: Numbers in *italics* indicate figures and numbers in **bold** indicate tables on the corresponding page.

#*LancsBox* 226, 391, 400

AAC and Non-AAC User Workplace Corpus (ANAWC) 268

Aarts, J. 540

aboutness 541

action ascriptions 256

action description 245

action formation 245

active voicing 519

Acton, H. 245

Adami, E. 574

adjectival modifiers 342, 342

Adolphs, S. 8, 15, 20, 70, 113, 117

affirmative and negative patterns **155, 156**

affordances 203–204, 205

aggression space 113

Aijmer, K. 13, 16, 109, 112, 117, 151

Aitchison, J. 145

Alexander, M. 307, 333

Alfano, M. 453n13

algorithms 572

alignment 75, 77

Allington, D. 213, 534

Alvarado, R. 186, 196

Amager project 189

American and British Office Talk Corpus (ABOT) 267–268

AMI Meeting Corpus 71, 73

Amsterdam Metaphor Lab 130, 139

Ancient Greek and Latin Dependency Treebank (AGLDT) 409, 411–412

Ancient Lives 533, 546–551, 547, 548, 549, 550, 551

Andersen, G. 109, 150, 151

Anderson, C. 213

Anderson, S. 560

Andow, J. 2, 444, 446, 449, 453n16

Androutsopoulos, J. 289, 574

annotation: digital pragmatics and 114; of multimodal corpora 72–73, 75, 92; versus non-annotation 383; of speech acts 116; of spoken corpora 10–11, 67; in stylistic analysis 312, 313; tools for 80

Anstey, F. 318

AntConc 226, 240n17, 266, 336, 539

Anthony, L. 235, 310

anti-racism 50

Antonijević, S. 479–480, 485

Anvil 75

apologies 113, 115

apophenia 189

applied linguistics 15–16

Archer, D. 113

A Representative Corpus of Historical English Registers (ARCHER) 31

asterisks 50

Atanasova, D. 518

ATLAS.ti 54

attitudinal meanings 68

ATT-Meta Project Databank 130

audio/visual channels 288

auditory approach 69

Austin, J.L. 108

Australian National Dictionary (AND) Wordnik 174

authenticity 194, 195

author–recipient relationship 44

authorship analysis 360, 361, 363, 364–365,

373, 495

AuToBI 75, 76, 77, 78

autocorrelation 347–348

auto-ethnographic approach 190

autoregressive integrated moving average

(ARIMA) 349, 350

auxiliaries, passive 340, 341

axiology 204, 210–213

- backchannels 21
 Baker, C. 512
 Baker, H. 120
 Baker, P. 20, 171, 222, 235, 378, 379, 386, 387, 388, 390, 394, 516
 Bakhtin, M. 466, 575
 Bakhtin, M.M. 517
 Baldry, A. 90, 92
 Balossi, G. 313
 Bamman, D. 291, 299
Bank of English (BoE) 7, 27, 127, 148, 382
 Baraibar, A. 478
 Barnden, J. 130
 Barthes, R. 87
BASE Corpus (British Academic Spoken English Corpus) 16
 Bateman, J. 90, 92
 Bates, V. 514
 Bauman, R. 206
BE06 381, 392
 beat gestures 70, 79
 Becker, K. F. 146
 Bednarek, M. 517, 519
 Beeching, K. 14
 Belliger, A. 207, 214
 Bell, P. 93, 94
 Benardou, A. 559
 Bentham, J. 5, 215
Bergen Corpus of London Teenage Language (COLT) 8, 14, 109
Berkeley Master Metaphor list 129
 Berry, D.M. 484
 Best, S. 311
 Biber, D. 8, 19, 20, 31, 41, 149, 158, 265, 269, 332, 379
 Bible 131, 144
 big data: digital pragmatics and 113; domination of 54; health humanities and 516; historical linguistics and 331; mediated discourse analysis and 213; patterns and 189; sociolinguistics and 293–294; spoken corpora and 10
Big Tent Digital Humanities 456
Birmingham Blog Corpus 120
Blackwell Companion to the Digital Humanities 556
Bleak House (Dickens) 314–318, 316, 317
 Blommaert, J. 198
Bloomsbury Companion to Stylistics, The (Sotirova) 307
BNC-Baby corpus 130
 Bødker, M. 190
 Boersma, P. 69
 Bohmann, A. 294, **295**
 Bolden, G.B. 250
 Boon, S. 51
 Borgman, C. 288
 born-digital data 66, 73, 483
 Bortolussi, M. 499
 Bourhis, R.Y. 275
 boyd, d. 189, 197, 293
 Boyer, E.L. 478
 Brexit 223
 Brezina, V. 315
 Briggs, C.L. 206
British Academic Spoken English Corpus (BASE) 7
British Academic Written English (BAWE) 34, 35–36, **36**
British National Corpus (BNC) 12, 27, 34–35, 71, 127, 148, 154, 170, 238, **321**, 368, 378, 381; *see also Spoken British National Corpus (BNC)*.
British Telecom Correspondence Corpus (BTCC) 39
 Bronson, B. 29
 Brookes, G. 394, 516
 Brown, B. 521, **522**
Brown Corpus of American English 27, 29, 285, 299n2, 338–340, 339, 382
 Brown, P. 111
 Brown, R. 154
 Bullokar, W. 144
 Burdick, A. 86
 Burke, M. 307, 507n10
 Burt, M.K. 154
 Busa, R. 29
 Butler, J. 208
 Button, G. 256
 Cairo Genizah 407, 414n9
 Caldas-Coulthard, C.R. 31
 Cambridge Analytica 215
Cambridge and Nottingham Business English Corpus (CANBEC) 267
Cambridge and Nottingham Corpus of Discourse in English (CANCODE) 7, 12, 16, 113
Cambridge and Nottingham e-language corpus (CANELC) 380
Cambridge Grammar of English 19
Cambridge Grammar of Spoken and Written English, The (Carter and McCarthy) 149
Cambridge Handbook of Stylistics, The (Stockwell and Whiteley) 307
Cambridge International Corpus (CIC) 149
Cambridge Learner Corpus (CLC) 153, 154
 Caple, H. 89
 Carletti, L. 513
 Carmichael, A. 212
 Carrol, G. 503
 Carter, R. 8, 16, 20, 149, 319
 Carter, R.A. 143, 158
 cartoons 457, 458

- Case, D. 558
 Cayley, J. 206
 Cenoz, J. 276
 Chamberlain, A. 190
 Champion, E.M. 164
 Chang, H. 464
Changing the Education Paradigm (RSAnimate) 464, 465
Charles Dickens Corpus 32
 Charmaz, K. 56
 chauvinism 450–452
 Chen, Y. 70
 Child ballads 29
Child Language Data Exchange System (CHILDES) 10
 Chilton, K. 150
 Chiluba, I. 50
 Chinese 50
 Chomsky, N. 283, 443
 Chovanec, J. 221
 Christian theology 131
 Christie, A. 499
Christmas Carol, A (Dickens) 319
 Christopherson, S. 208, 209
Chronicle of Higher Education 475
 citizen science: background 531–534;
 communities of practice 535–536; computer-mediated communication and 537–538;
 corpus-based versus corpus-driven 539–540;
 corpus linguistics and 538–541; critical issues and topics 534–538; crowdsourcing 534–535; current contributions and research 541–546; definition of 530; diachronic dimension 536–537; future directions 551–552; keywords 540–541; lexical features 544–546; lexical selection 537; as model 533–534, 533; research methods 538–541; sample analysis 546–551; Zooniverse *see* Zooniverse;
CLAN 10, 258n7
 Claridge, C. 112
 Clark, C. 323
 classics: background 408–409; critical issues and topics 411–412; current contributions and research 409–411; future directions 412–413; interaction of languages 405–408
CLiC Activity Book (Mahlberg) 323
CLiC Dickens 32, 496
CLiC web 311, 313, 315, 317
 CLMETEV 331
 close reading 189
 Clough, P.T. 516
 clusters 20, 33, 182, 290, 312, 350–351, 540
CMU Sphinx 75
 Cobbett, W. 145
 Cochrane, D. 519
 coding: of animal and human behaviour 75;
 digital literacies and 483–485; focused 56; in gesture research 70; of online live data 75
 Cohen, S. 478
 Cohen, W.W. 369
 colligation 171
Collins Birmingham University International Language Database (COBUILD) 30, 33, 147–148
 Collins, M. 570
 collocation 26–27, 128, 171, 393–398, **395**, **396**, **397**, 539
 collostructional analysis 350
 collustration 238
Cologne Dialogue on Digital Humanities 2012, The 560
 comics 458, 461
 commissives 111
Common European Framework of Reference for Languages (CEFR) 153
Common Languages Resources and Technology Infrastructure (CLARIN) 561
 communicative pragmatics 276–278
 communities of practice 535–536
Companion to Digital Humanities (Hockey) 309
 competency focus 153, 157
 compliments 110, 111, 114, 116
 computational linguistics 290–291, 298–299
 computational literacy 484
 computational sociolinguistics 290–291
 computational turn 197
 computer-assisted analysis 312
Computer-Assisted Qualitative Data Analysis Software (CAQDAS) 89
Computer Assisted Textual Markup and Analysis (CATMA) 309
 computer-generated attribution 495
Computerized Language Analysis (CLAN) 7, 250
 computer-mediated communication 289, 537–538
 computer-mediated discourse (CMD): definition of 289; forensic linguistics and 366–367; methodologies in 362; orthographic standards and 287; sociolinguistics and 292; whiteboard animation 456–457
 conceptual analysis 441–442, 446
 conceptual metaphor 125, 127, 129
 concordances: computerised, development of 29; corpus linguistic use of 311, 398–399; definition of 539; examples of **315**, **319**, **320**, **321**, **322**, **398**, 424; KWICGrouping and 313; uses of 128, 311; as visualisation 311
CoNe: A corpus of Narrative etymologies 333
 Conklin, K., 503
 connotative meaning 87
 Conrad, S. 20

- consent 52
 content selection 561
 context: axiology and 211; conversation analysis and 246–247; corpus linguistics and 380; digital pragmatics and 113, 114; ethnography and 188, 198; lexis and 170–171
 contextualisation as part of the previous 214
 Continental European pragmatics 107
 contrastive analysis 114
 contrastive focus 151
 contrastive interlanguage analysis (CIA) 152
 convergence culture 456
 conversation analysis (CA): action ascriptions 245–246, 247; automation of 256; background 243–244; context and 246–247; critical issues and topics 245–247; current contributions and research 251–255; data collection 247–248; digitalisation of 250–251, 251–252; focus of 19, 243; future directions 255–256; methodological developments and challenges 252–255; of online interaction 253–255; research methods 247–251; systematicity in 150; transcription 248–249, 250–251, 254–255
 Cook, G. 150
 co-production 562–563
 Cop, U. 499
 Corbett, J. 498
 Cord 333
 Cordell, R. 561
 Corder, S.P. 151
 CORE (COnnecting REpositories project) 560
 corpora: cross-cultural communication and 264–265; cross-cultural workplaces and 266–270; definition of 26, 264, 378, 539; general 378; grammar and 153–154; health language 516; historical linguistics 329–333; influence on grammar 146–150; learner 16–17, **28**, 154; newspaper 419–420; parsed 330; reference 170; as relevant population data 368–371; research approaches 265–266; specialised 20–21, **28**, 368, 378; spoken *see* spoken corpora; types of **28**; written *see* written corpora;
 corpus-assisted critical discourse analysis 222
 corpus-assisted discourse analysis 20–21
 corpus-based methodologies: versus corpus-driven 539–540; digital pragmatics and 113, *115*; keywords and 518; lexis and 173; metaphor and 127–129; multimodality and 86–87, 92, 382; philosophy and 449–450, 452
 corpus-based speech act retrieval *115*
 corpus-driven methodologies 173, 329, 382, 540
 corpus linguistics: annotated versus non-annotated 383; background 379–380, 531; challenges of 238; citizen science and 538–541; collocation 393–398, **395**, **396**, **397**; concordances 398–399, **398**; corpus-based versus corpus-driven 382; critical discourse analysis and 222, 384; critical issues and topics 380–385; cross-cultural communication and 264–265; culturomics and 386, 389–390; current contributions and research 385–390; data collection 382–383; definition of 378, 539; described 264–265; digital humanities and 385–386; early use of 108–109; English language studies and 570; frequencies 391–392, **391**; future directions 400–401; keywords 392–393, **393**; limitations of 379–380; mode of communication 380–381; multilingual versus monolingual corpora 384–385; objectives of 5; public humanities and 538–541; research approaches 265–266; sample analysis 390–399; strength of 19; total accountability versus data selection 383–384
Corpus of American Soap Operas 18
Corpus of Contemporary American English (COCA) 12, 27, 36–37, 37, **38**, 154, 170, 331, 368, 383, 387
Corpus of Early English Correspondence (CEEC) 31
Corpus of Early English Medical Writing 332
Corpus of English as Lingua Franca in Academic settings 17
Corpus of English Dialogues (CED) 18, 117
Corpus of English Novels (CEN) 32
Corpus of English Religious Prose 31
Corpus of Global Web-based English (GloWbE) 285
Corpus of Historical American English (COHA) 31, 37–38, 172
Corpus of Historical English Law Reports 1535–1999 (CHELAR) 31
Corpus of International Learner English (ICLE Corpus) 16
 corpus pragmatics 109
Corpus Query Language (CQL) 42, 154
Corpus Query Processor (CQP) 7
 corpus stylistics: *see* literary stylistics.
 correlational and distributional relationships 251–252
 correlation metaphors 137
 correlograms 348
 Cortes, V. 269
 Coulthard, M. 368, 373
 Council of Europe 153
Council of European Social Science Data Archives (CESSDA) 561
 Couper-Kuhlen, E. 255
 Cousins, J. 210
 co-verbal gestures 69, 70

- Coxhead, A. 30
CQPweb 180, 226
 Crabtree, A. 190
 Crane, G., 556
 Crawford, K. 189, 197, 293, 524
 Crawford, P. 512
Creative Language System Group 129
 Creative Practice as Mutual Recovery 517
 Creese, A. 188
 Criado-Perez, C. 131
 crime, cyber-enabled 362–363
 Criminal Practice Directions Amendment No.2 361–362
 critical discourse analysis (CDA): analysis 226–238; approaches to 150, 222; background 222–225; concordance table **227**; corpus linguistics and 384; data collection 223–225, 225; definition of 220, 222; future directions 238–239; keywords **227**; pronoun use 232–234, **233**; prototypical texts 234–238, **235**; Romanians 228–232, **230**
 critical literacy 485
 cross-cultural communication: corpora and 264–265, 266–270; corpus linguistics and 264–265; current contributions and research 265–270; digital photography and 274–278; sample analysis 270–274
 cross-cultural workplaces and 266–270
 cross-linguistic comparative studies 251
 crowdsourcing 534–535, 551; heritage 513
 Crystal, D. 68, 147
 Culkin, J.M. 204
 Culpeper, J. 18, 113, 313
 cultural heritage: background 555–557; conclusions 563–564; co-production 562–563; critical issues and topics 557–560; digital content used by scholars 558–559; emerging research and key analysis 561–563; environments, creating 562; future directions 560–561; impact of digital technology on 557–558; tools for 559–560
 cultural tools 203–205
Cultures of Knowledge 31
 culturomics 386, 389–390
 Cunningham, R. 480
Cut, The (Taylor) 193, 194
 Cuyckens, H. 336
 Cyclic Gesture 70
Cymru1914.org 562–563
 Czech language 408

D64 corpus 74
 Dagenais, D. 275
Daily Express, The 224
 Danet, B. 50, 569, 573
 ‘Dark Matter’ corpus 544–546

Darwin Correspondence Project 31
 data: as epistemological alibi 208; as lexical issue 171–173; size of 71–72
 databases 40, 220
 data collection: in conversation analysis (CA) 247–248; in corpus linguistics 382–383; in critical discourse analysis (CDA) 225; in critical in discourse analysis (CDA) 223–225; in forensic linguistics 369; in historical linguistics 333–334
 data literacy 488
 datalogical turn 516
 data mining 189
 data selection approach 383–384
 Daubert Criteria 362, 363
 Davidson, C.N. 475, 481, 487
 Davies, B. 56
 Davies, M. 113, 291
 Day, G. 514
Day in the Life of the Digital Humanities 468
 decontextualisation 189, 206
 deictic gestures 70
 Deignan, A. 127
 Dekker, R. 484
 Deleuze, G. 208
 Demjén, Z. 131
 Denis, D. 291
 denotative meaning 87
 dependency grammar 407, 408
 Descartes, R. 208
 design ethnography 190
 Deumert, A. 50
Deutsches Wörterbuch 329
 Deutschmann, M. 115
DHW project 559
Diachronic Corpus of Present-Day Spoken English (DCPSE) 13
 diachronic dimension 536–537
 diachronic/historical corpora 31–33, 108
 diachronic speech act analysis 114
 diachronic variation 112
 diachrony 347
 dialectal corpora 10
 dialectometry 291
 dialogue, genres of 11
 Dickens, C. 503; *Bleak House* 314, 316; *Christmas Carol, A* 319; *CLiC Dickens* 32, 496; clusters in 496; *Great Expectations* 319–320
 Dickey Fuller test 349
Dictionary of American English on Historical Principles, A (DAE) 174
Dictionary of American Regional English (DARE) 174
Dictionary of Old English (DOE) 174
Dictionary of the Scots Language (DSL) 174
 Diez-Bedmar, M.B. 152

- Digging into Connected Repositories*
(DiggiCORE) 560
- digital data case study 294–298, **295**, 297
- Digital Editions for Corpus Linguistics*
(DECL) 44
- digital ethnography: *see* ethnography.
- digital health humanities: background 512–513;
critical issues and topics 513–515; current
contributions and research 515–517;
digitised language 517–523; future directions
523–524; keywords in 521, **521**; sample
analysis 517–523
- digital humanities: benefits of 556–557; centres
556; corpus linguistics and 385–386; corpus
stylistics and 312; critical issues and topics
110–111; cultural heritage and 556; definition
of 2, 5, 86, 107, 186, 242, 263, 456, 468, 479,
511, 556, 570; digital literacies and 478–480;
early developments 2; economic agendas
and 213; English and 126; English language
studies and 570, 571; ethical research in
52–53; ethnography and 185; expansion of
516; health humanities and 524; historical
linguistics and 328–329; ideological agenda
of 213; interdisciplinary research and
418–419; literary stylistics and 309–310,
310–311, 494; metaphor and 125; nature of
570; objectives of 5; second wave 534; tools
for 309–310; use of term 38–39
- Digital Humanities Quarterly* 480
- digital inclusion 488
- digital interaction: critical issues 52–53; current
contributions and research 50–51; definition
of 49; future directions 62; Mumsnet study
53–62
- digital literacies: acquisition of 485; background
474–475; born-digital content and 483; in the
classroom 479; coding 483–485; components
of 481; conditions for 475–481; content
475–476; critical issues and topics 475–481;
critical literacy 485; current contributions
and research 481–482; definitions 473–474;
digital humanities 478–480; digital inclusion
and 488; digital scholarship and 477–478;
digitised surrogates and 482–483; formal
courses 486; future directions 488–489;
general 481–482; identities 477; learning
environments and 481; literacy frameworks
486–487; pedagogical issues 487; practices
476; research methods 482–487; self-
taught 485–486; visual literacy 476–477;
workshops 486
- digital photography 274–278
- digital pragmatics: annotation 114; background
108–109; current contributions and research
112–116; definition of 107; future directions
120; research methods 113–116; sample
analysis 116–120; *see also* pragmatics.
- Digital Replay System (DRS)* 75
- Digital Research Infrastructure for the Arts and
Humanities (DARIAH-EU)* 556, 560
- Digital Research Tools (DiRT)* 559
- digital scholarship 477–478
- Digital Scholarship in the Humanities* 480
- Digital Scholar: The* (Weller) 477–478
- Digital Skills: Unlocking the Information
Society* (van Deursen and van Dijk) 475
- digital technologies: conversation analysis and
250–251, 251–252; evolution of 1–2; history
and 418, 419; impact on the humanities
557–558; and learning 472; limitations of
89–90; psycholinguistic 500–501, 506;
scholarly use of 558–559
- digital turn: cultural heritage and 556; in
English language research 1–2; grammar
and 146–147; historical linguistics and 329;
in the humanities 1–2; metaphor and 125;
multimodal narrative and 458; sociolinguistics
and 283, 285–298
- dimension scores 41, **41**
- directives 111
- discourse: computer-mediated 287, 289, 292;
definition of 289; digital 286; digitally
mediated 289; powerful forces as 53;
sociolinguistics and 289; spoken corpora and
19–21
- discourse markers 19
- distant reading 189, 209, 311
- Distributed Red Hen Lab* 130
- Diver* 75
- diversity in samples 450–451
- Dixon, P. 499
- 'Do Artifacts Have Politics?'* (Winner) 212
- Doceri 461
- domains WorldBuilder 313
- domain words 541
- Donaldson, C. 422
- double-voicing 466
- Drauker, F. 575
- Drew, P. 150
- Drucker, J. 288
- Dulay, H.C. 154
- Dumitrica, D. 195
- dysfluency phenomena 19
- e-AQUA Project* 413
- Early English Books Online (EEBO)* 113, 120,
173, 175, 176, 332
- Early Modern English Books Online* 385
- Early-Modern Multiloquent Authors (EMMA)*
332, 333
- ecological validity 500–501

- Economics Is for Everyone!* (RSAnimate) 464–465, 465
- Eder, M. 310
- Educational Testing Service (ETS)* 33
- Education for Change LTD* 559
- EEBOCorp* 331–332; *see also* Early English Books Online (EEBO).
- Ehrlich, S. 250
- Einstein, A. 502
- eLALME* 333
- ELAN* 10, 75, 76, 77, 250, 258n10, 266, 290
- El Bouanani, S.E.M. 365
- El Corpus de Lenguaje Adolescente de Madrid (COLAm)* 12
- electrodermal activity (EDA) 500, 506
- Electronic Enlightenment* 31
- ellipsis 149
- Ellis, N.C. 154
- embodied action 250
- Emmott, C. 307
- emoji 50, 292, 579–581
- encoding 40
- English as a lingua franca 17, 541–543
- English Dialect Dictionary (EDD)* 174
- Englishes 13, 284
- English Grammar, Adapted to the different classes of learners* (Murray) 145
- English Grammar Profile* 153, 156
- English language and classics: *see* classics.
- English language and digital cultural heritage: *see* digital cultural heritage.
- English language and digital health humanities: *see* digital health humanities.
- English language and digital literacies: *see* digital literacies.
- English language and English literature: *see* English literature.
- English language and history: *see* history.
- English language and multimodal narrative: *see* multimodal narrative.
- English language and philosophy: *see* philosophy.
- English language and public humanities: *see* public humanities.
- English language and social media: background 569–571; corpus linguistics and 570; creative spellings 577; critical issues and topics 571–572; current contributions and research 572–575; definition of 569; digital humanities and 570, 571; emoji 579–581; future directions 581–582; images 577; multimodality and 574–575; sample analysis 575–581
- English language teaching materials 33–34
- English literature: *see also* literary stylistics.; background 494–498; corpus tools for 495; current contributions and research 498–503; future directions 505–506; literary stylistics and 496–497, 498; psycholinguistics and 494–495, 498–502; sample analysis 503–505
- Enron case study 368–371
- entextualisation 206, 214
- ENUMERATE* 555
- epistemology 204, 208–210, 447
- Erman, B. 13–14
- Eslambolchilar, P. 190
- ethics 52–53, 224, 240n12, 364
- ethnography: approaches to 188, 189–190; background 187–188; context and 188; critical issues and topics 188–189; current contributions and research 189–190; digital humanities and 185; future directions 196–198; labels for 187; online/offline distinction 188, 191; research methods 190–191; sample analysis 191–196
- ethnomethodological approach 190
- ethnomethodological conversation analysis 257n1
- ethnomining 189
- Europeana* 555
- Europeana Cloud: Unlocking Europe's Research via The Cloud* 557, 558, 560, 562
- event-related potentials (ERPs) 501, 506
- evidential verbs 44
- Ewert, J.C. 461
- e-Wissenschaft* 556
- expanding circle 542
- experiential data 498–499
- experimental philosophy 445–446, 449–450, 453n8, 453n10, 453n11
- expressives 111, 114, 115
- Extensible Markup Language (XML) 10
- eye tracking 501–502
- Faber, P. 137
- Facebook* 292
- facial electromyography (EMG) 500, 506
- factor analysis 295–296, 297
- Fahlman, S. E. 573
- Fairclough, N. 222
- Falconer, L. 132
- Familjeliv 51
- Färdigh, M.A. 221
- Fauconnier, G. 125
- FAVE-align interface* 76, 77, 290
- F-Brown corpus* 338–340, 339
- Federalist Papers* 364
- Ferreira-Junior, F. 154
- Fialho, O. 307, 323
- fictional speech 312–313
- figurative language 129; *see also* metaphor.
- Finegan, E. 31

- Firth, A. 543
 Firth, J.R. 29, 171, 394, 422
 Fitzpatrick, K. 571
 Fletcher, E.H. 516
 fMRI 501, 506
 folklore 29
 forced alignment 77, 80
 foregrounding 307
 forensic linguistics: assumed identities online 371–373; authorship analysis 360, 361, 363, 364–365, 373; background 360; computer-mediated discourse and 366–367; critical issues and topics 361–364; current contributions and research 367; definition of 360; digital methods, need for 362–363; Enron case study 368–371; future directions 373; linguistic identity 366–367; reliability of evidence 363–364; research methods 364–367; sample analyses 368–373; stylistic approaches 365–366; stylometry 364–365
 form-to-function approach 110
 Foucault, M. 53
 Fox, B.A. 252
 Fozdar, F. 50
 Francis, G. 33
 Frank, M. 488
Freiburg-Brown corpus of American English (Frown) 382
Freiburg-LOB Corpus of British English (FLOB) 382
 Freire, P. 473
 frequencies: corpus linguistics and 391–392, **391**; definition of 539; in digital health humanities 521; in English literature research 495, 496; in historical linguistics 335, 336; interpretation of 389; lexis and 169–170; quantitative analysis of 337–338
 Fries, C.C. 29
 Fries, P.H. 29
 Friginal, E. 265, 270
 Fröhlich, D.M. 256
 Fuchs, C. 197
 Fuchs, R. 291
 function-to-form approach 110
 Fung, L. 16

 Gablasova, D. 158, 394
 Gabrielatos, C. 171
 Gaden, G. 195
 Galleron, I. 310
 Garcia, A.C. 576
 Gee, J. 475
 Gee, J.P. 203, 205, 487
 GeM framework 90, 92
Gendered Conference Campaign 453n20
 gendered nouns 387–390, 387
 gender inequity 466
 gender-specific differences 14
 generalised linear mixed-models 346–347
 Genette, G. 496
 genitive constructions 338–340, 339, 345, 346
 genres 11–12
 geographical text analysis (GTA) 419, 422–424, 433
 geoparsing 422, 436
 Georgakopoulou, A. 289, 459
 Gerbner, G. 220
 Gershon, I. 191, 192
 gesture studies 69–70, 74, 75, 77–79, 79, 381
 Ghesquière, L. 342
 Gibbs, F. 242
 Gibbs, R.W. Jr. 127
 Gibson, J.J. 203, 205
 Gilbert, P. 515
 Giles, D. 253, 254
 Gillen, J. 487
 Gilquin, G., 152
 Gilster, P. 474
Girl on the Train, The (Hawkins) 308
 Gledhill, C. 538, 541
Global Web-Based English corpus 368
GloWbE 38–39
 Goatly, A. 129
 Goffman, E. 194, 244, 248, 252, 289, 575, 577
 Goldberg, D.T. 480
 Gong, Y. 51
 Gonzalez-Lloret, M. 254
 Gorter, D. 275, 276
 Grady, J.E. 137
 grammar: audio recordings and 146, 147; background 144–150; corpora and 146–150, 153–154; critical issues and topics 150–153; current contributions and research 153–154; definition of 143; error or competency focus 151–153, 157; future directions 158; historical development of 144–148; learner 151–153; sample analysis 154–157; second language and 153–154, **155**, **156**; spoken 8–9, 19, 149; vs. structure 143; systemic-functional 87; visual 87
Grammar of Spoken English, A (Smith) 146
Grammar of the English Language, In a Series of Letters, A (Cobbett) 145
 Granger, S. 152
GraphColl 315
 graphic novels 458
 gratitude expressions 116–120
Great Expectations (Dickens) 319–320
 Greek 406–407, 411, 413n5
 Greenhall, A. 208, 209
 Greetham, D.C. 497
 Gregory, I. 422
 Grice, P.H. 108
 Gries, S.T. 334, 335

- Grieve, J. 570
 Griffiths, R. 558
 Groom, N. 538, 541
 Guattari, F. 208
 Guy, J., 503
- Habermas, J. 222
 Hacking, I. 540
 Hadikin, G. 537, 538
 Hafner, C.A. 481, 482, 485
 Hakemulder, F. 494, 495, 496
 Hall, G. 312, 557
 Halliday, M.A.K. 42, 87, 88
 Hall, M. 51
 Halpern, O. 206, 210
 Hamilton, W. 546
HamleDT normalisation tool 409
 Handford, M. 518
Hansard Corpus 38, 38, 180, 181, 181, 332, 347, 348, 349, 350
 Hansen, N. 453n14
 Happy Land 51
 Hardaker, C. 131
 Hardie, A. 379, 380, 381, 382, 384, 400, 401
 Harré, R. 56
 Harvey, K. 521, **522**
HASTAC (Humanities, Arts, Science, and Technology Alliance and Collaboratory) 480
 Hawkins, P. 308
 Hayles, N.K. 205
 head nods 21, 71, 79
 healthcare communications 15, 269–270
 health humanities 511, 514, 515, 516, 524; *see also* digital health humanities.
Health Humanities (Crawford et al.) 514, 515
Health Language Research Group 516
 Heath, C. 251
 Hebrew 407
 Heino, R.D., 131
 Heller, M. 284
Hellespont Project 412
Helsinki Corpus of Historical English Texts (HC) 329–330
Helsinki Corpus of Older Scots 332
Helsinki ELFA Corpus 17, 109, 113, 329
 Henty, M. 558, 559
 Hepburn, A. 250
 Heritage, J. 252
 Herring, S. 367, 372
 Herring, S.C. 289, 569, 573
 Hesson, A. 269
Hexapla 407, 413
 Highfield, T. 196
 Hilpert, M. 334, 335, 338
 Hine, C. 187, 190
 Hinrichs, L. 338
 Hintikka, J. 443
 historical/diachronic corpora **28**
 historical linguistics: background 328; corpora 329–333; critical issues and topics 328–329; data analysis 336–352; data retrieval and preparation 333–334; digital humanities approach 329; future directions 353; parsed corpora 330; research methods 329–352; responsible use of sources 334–336
Historical Thesaurus 135–138, 176, 177, 178, 180
Historical Thesaurus of English (HTE) 117, 126, 135, 174–175
Historical Thesaurus Semantic Tagger (HTST) 180
 history: background 419–420; conclusions 433–436; digitisation and 418; issues and research in the study of poverty 420–421; research methods 422–424; sample analysis 424–433
 Hitchcock, T. 561
 Hobbs, F.D.R. 269
 Hockey, S. 309, 310
 Hoey, M. 318
 Hoffmann, S. 333
 Hofstede, G. 99
 Holmes, J. 114
 Holocaust 560
 homonymy 172
Hong Kong Corpus of Spoken English (HKCSE) 9, 72, 267
 Honkapohja, A. 44
 Hornby, A.S. 33
 Houghton, J. 558, 559
 Howatt, A.P.R. 145
How to Help Every Child Reach Their Potential (RSAnimate) 467
HuComTech 73
 Huey, E.B. 505, 506
 humanities: computing in 1, 186, 309; definition of 5, 413n1; in pre-loaded corpora 34–39, **36**, 37; as semantic shifter 34; so-called 39
 Hume, D.E. 207
 Hunston, S. 29, 33, 146, 538
 Hutton inquiry 20
 Hyland, K. 34
 HyperResearch 89
 Hypertext Markup Language (HTML) 96
 Hypertext Preprocessor (PHP) 96
 hypothetical values **94**
- IBM/Lancaster SEC* 72
ICAMET-letters 332
iConc 267
 idioms 168
 Iedema, R. 208, 209
 illocutionary force 111, 115
 illustrative eclecticism 120

- image-act 87
 Informatics Laboratory 130
 instrumental approach 69
 insults 119
Intellectual Humility and Cultural Diversity in Philosophy 453n19
INTERACT 75
 interlanguage 151–152
International Association of Forensic Linguists (IAFL) 362
International Corpus of English (ICE) 8, 11, 152, 285, 294, **295**
International Corpus of Learner English (ICLE) 152
 Internet: access to 474; as data source 73; identities and 477; paradigm shift and 475
 Internet Relay Chat 50, 292
Intonational Variation in English (IViE) 72
 intuition 442–443, 444–445, 449, 453n10
 Irigaray, L. 208
Irish Emigration Database (IED) 32
 Isidore 560
 ITHAKA Report 558

 Jacobs, A.M. 499, 506n3
 Jacobs, J.B. 576
 Jacobsson, M. 118
 Jaffe, A. 284
 James, H. 502, 503
 Jautz, S. 117
 JavaScript 96, 485
 Jaworska, S. 51
 Jefferson, G. 244, 248
 Jenkins, H. 456, 458, 476
 Jensen, K.E. 2, 3, 222, 310, 385
 Jespersen, O. 146
 Jewitt, C. 92
 Jisc 481
 Jockers, M.L. 309, 311
 Johansson Falck, M. 127
 Johnson, A. 369, 370
 Johnson, M. 129
 Johnson, S. 44
 Johnstone, B. 284
 Johnston, T. 381
 Jones, R. 572
 Jones, R.H. 481, 485
 Jonsson, C. 574
Journal of Open Humanities Data 288
 JSTOR 444
 Jucker, A.H. 109, 114, 115, 116
Julius 75
JUXTA 495

 Kachru, B.B. 541–542
 Kärkkäinen, E. 79
 Kassou, I. 365

 Katsuno, H. 50
 Kauhanen, H. 147
 Kawara, O. 202
 Kay, C. 333
 Kehoe, A. 120
 Kelly, David 20
 Kemp, S.J. 514
 Kendon, A. 69, 70
 keywords: analysis of 226, **227**, 392–393, 518, 540–541; definition of 392, 518, 521, 540; digital health humanities and **521**; frequencies and 392–393, **393**; place-name co-occurrence and **431**
 Kirk, J.M. 150, 151
 Kirschenbaum, M. 86, 129, 197, 263, 484, 556
 Kirschenbaum, M.G. 126, 309, 479, 524
 Kirwan, P. 507n7
 Klee, M. 205–206, 207, 212
 Knight, D. 381
 Knobel, M. 473, 475
 knowledge, production and diffusion of 531–532, 531, 533
 Knox, J. 89
 Kohnen, T. 116
 Koltay, T. 511
 Koppel, M. 365
 Kövecses, Z. 127, 135
 Kramsch, C. 277
 Kredens, K. 368, 373
 Kress, G. 87, 88, 95
 Krieger, D.J. 207, 214
 Kruisinga, E. 146
 Kuiken, D. 498, 499, 507n8
KWICGrouper 315, 316
 Kytö, M. 18, 112

 Labov, W. 284, 459
 Ladewig, S.H. 70
 Lado, R. 151
 Ladyman, J. 444, 445, 446
LAEME 332
 Laing, M. 333
 Lakoff, G. 129
Lancaster/ IBM Spoken English Corpus (SEC) 9
Lancaster-Oslo/Bergen Corpus (LOB) 27, 29, 382
LancsBox 315
 Landry, R. 275
 language awareness 275–276
 language ideology 191
Language@Internet 289
Language in the Workplace Project (LWP) 267
 language use, specific and general 319–322
 Lankshear, C. 473, 475
 Larsen-Freeman, D. 143, 157
 Larsen, J. 190
 Lass, R. 333

- Latin 144–145, 406–407, 411, 413n4
 Latour, B. 202, 206, 212
 law: *see* forensic linguistics.
 learner corpora 16–17, **28**, 154
 learning environments 481
 Leech, G. 146, 148, 150, 307, 310, 383, 385, 391
 Lee, C.K.M. 50
 Lee, J.J. 311
 Lefebvre, H. 277
 Lehtonen, S. 51
 Leinweber, D. 189
 Lemke, J.L. 89
 letters 31–32, 39–44
Letters of Artisans and the Labouring Poor (LALP) 39
 Levinson, S.C. 108, 111
 Levshina, N. 338
 lexical affiliates 70
 lexical bundles 20, 33, 34
 lexical innovation 570–571
 lexical selection 537
 lexis: background 165–167; challenges of 169–171; critical issues and topics 167–173; current contributions and research 173–176; data issues in 171–173; future directions 181–182; importance of 164; resources 173–175; sample analysis 176–181; tools 175–176
 Liddicoat, A. 244
 Lightfoot, D. 329
 limitations of digital tools 89–90
 linear data 89–90
lingua franca, English as 17, 541–542, 569
Linguistic DNA of Modern Thought (LDNA) 176, 182
 linguistic identity 366–367
Linguistic Innovator Corpus (LIC) 14
 linguistic landscape 263, 274, 275–278
 linguistic variation 283–284, 292, 294–298
 Linked Reading 311
 Linn, A.R. 144, 145, 147
 Lin, P. 71, 73
LION 506n6
 listener responses 21
 literacy frameworks 486–487
 literary linguistics: *see* literary stylistics.
 literary processing 499
 literary stylistics: background 306–307; critical issues and topics 307–310; current contributions and research 310–331; definition of 306, 307; digital humanities and 310–311, 312, 494; foundational facts 308; future directions 322–323; language use 319–322; linguistic analysis and 498; meaning through repetition 313–318; patterns in narrative 318–319; research methods 312–313; sample analyses 313–322; text editing and 496–497
 literature studies 29
 Liu, A. 557
 Loeb classical library 414n6
 Loenneker-Rodman, B. 130
 log-likelihood 238, 240n17
London-Lund Corpus of Spoken English (LLC) 7, 11, 13, 71, 72, 109, 147, **147**
Longman Grammar Corpus 18
Longman Grammar of Spoken and Written English, The 33, 148
Longman Spoken and Written English Corpus (LSWE) 7, 19, 33, 148
Louvain International Database of Spoken English Interlanguage (LINDSEI) 16, 152
 Louwerse, M.M., 495
 Love Tracker 205–206
 Lowth, R. 145
 Luff, P. 251, 256
 Lutzky, U. 120

 machine learning 298, 299n4, 365
 macroanalysis 311
MacVisSta 89
 Madge, C. 51
Madness in Post-1945 British and American Fiction (Baker et al.) 512
 Mahlberg, M. 313, 317, 319, 541
 Malczewski, B. 15
 Malinowski, D. 277
 Mandala tool 309
 Manes, J. 114
 Manning, E. 33
 Manovich, L. 197
 Mapping Metaphor 135–138, 176, 178, *179*
Mapping the Republic of Letters 31
 Marcus, S. 311
 markup schemes 92, 94, **94**
 Marsh, J. 474
 Martinec, R. 89
 masculinity 51
Master Metaphor List 129
 Matthiessen, C.M.I.M. 42
 Mauraen, A. 15
 Mauro, A. 487
 MAXQDA 54
 McCarthy, K. 495
 McCarthy, M. 518
 McCarthy, M.J. 143, 144, 145, 149, 158
 McCloud, S. 458
 McEnery, A. 120
 McEnery, T. 220, 379, 380, 381, 382, 384, 386, 400, 401
 McGann, J.J. 496, 501, 503
 McGlashan, M. 238
 McIntyre, D. 313

- McLuhan, M. 204
 McNeill, D. 69
 mediated discourse analysis: aim of 214;
 definition of 203; further research 213–215;
 mediation *see* mediation;
 mediation: as axiology 204, 210–213; described
 203–205; as epistemology 204, 208–210; as
 ontology 204, 205–208
 medical humanities 511, 514
 Meho, L. 559
 Mendenhall, T.C. 309
 Mendoza-Denton, N. 288
 mental health 517–523
 mental processing 500
 Meredith, J. 253, 254, 255
 metadata 172–173
 metalanguage, theory of 34
 Metalude 129–130
 metaphilosophy 446
 metaphor: blood-related 128; body-part 128;
 cancer and 133–135; conceptual 125, 127,
 129; corpus-based approaches 127–129;
 correlation 137; critical issues and topics
 138–139; current research and methods
 126–132; in digital communication and
 environments 131–132; digital humanities
 and 125; digital resources and tools 129–130;
 future directions 138–139; journey as
 133–134; liquid 128; market-related 132; for
 money 178, 179; plant-related 127; polysemy
 and 172; reptile metaphors 135–138, 136,
 137, 138; resemblance 137; sample analysis
 132–138
 Metaphor Awareness 129
 metaphoric gestures 69–70
Metaphor in End-of-Life Care (MELC) 133, 139
 Metaphor Lab 130
 Metaphor Map 135–138, 136
 Metaphor Map of Old English 137
 Metaphor Menu 134
Metaphors We Live By (Lakoff and Johnson) 129
 metaphysics 447–448
 Meyer, M. 222
 Michel, J.-B. 312, 386, 387, 388, 390
Michigan Corpus of Academic Spoken English
 (MICASE) 7, 15
Microanalysis of Online Data (MOOD) 254
Middle English Dictionary (MED) 174
Migrant Correspondence Corpus (MCC)
 contains 39
 Millett, B. 44
Minorities and Philosophy 453n20
 modal verbs 155, 156, 157
 modes: communication 380–381; definition of
 67; gesture studies 69–70; speech prosody
 and 68–69; types of 67
 modifiers, adjectival 342, 342
MoMel 75
 monitor corpora 28, 382
 Monk Project 559
 monolingual corpora 384–385
 monologues 11
 Monterroza, C. 202–203, 204, 205
 Montfort, N. 484
 Montgomery, M. 517
Montreal Corpus 285
 Moon, R. 31
 Morel, E. 575
 Mosteller, F. 364
 Moya, J. Bednarek, M. 89
 Mroczek, K. 289
 Muhonen, A. 574
Multicultural London English (MLE) 14
 multidimensional analysis (MDA) 270–271,
 271, 272
Multidimensional Analysis Tagger (MAT) 41
 multilingual corpora 28, 384–385
Multilingual Internet, The (Danet and
 Herring) 569
 multilingualism 287–288
 multimodal concordance 91
 multimodal corpora 90–92
 multimodal discourse analysis (MDA) 88
 multimodality: *see also* multimodal narrative.;
 alignment 75; annotation 72–73, 75,
 92; approaches to 88; collation tools 75;
 critical issues and topics 70–73; current
 contributions and research 73–75; definition
 of 85; described 67; forms of 85; as future
 development 21, 22; future directions
 79–80; methodology 96; multimodal
 corpora 90–92; quality of recordings 72;
 research on 67–70, 85–86; sample analysis
 75–79; small data size and 71–72; in
 social media 574–575; social semiotic
 approach 87, 88; text and image *see* text and
 image; wearable devices 74; Web as data
 source 73–74; whiteboard animation and
 459–460
 multimodal narrative: background 457–458;
 critical issues and topics 458; current
 contributions and research 460–467; future
 directions 467–468; multimodality 459–460;
 narrative 459; point of view 462–465;
 research methods 459–460; setting 460–461;
 temporality 461–462; voice 466–467
 multi-word expressions (MWEs) 168
Mumsnet: absent fathers 57–60; background
 53; data construction 55–56; equal parenting
 60–62; gendered parenthood 57; identifying
 and analysing discourses 56; linguistic
 analysis 56; logo 54; mother as main parent
 57–60; qualitative approach 54; research
 design 55; as research site 51, 53; sample

- analysis 57–62; Your identity as mother 57–62
- Murakami, A. 154
- Murray, L. 145
- Murrieta-Flores, P., 422
- MySQL 96
- Mysteries and Problems of Reading, The* (Huey) 506
- Mysterious Affair at Styles, The* 499
- named entity recognition (NER) 422
- Narayanan, S. 130
- Narrative Corpus (NC)* 7
- native speakers 542, 543
- Nebraska project 412, 413
- NeDiMAH 560
- Neely, J. P. 210–211
- Network for Digital Methods in the Arts and Humanities (NeDiMAH)* 483
- Nevalainen, T. 331
- Newcastle Electronic Corpus of Tyneside English (NECTE)* 10
- New English Grammar, A* (Sweet) 146
- News on the Web Corpus (NOW)* 113
- newspaper corpora 419–420, 561
- newspapers, digital 113
- New Testament* 407
- nexus analysis 205, 214–215
- n-grams 33, 312, 561, 562
- NHS Comments Corpus (NHSCC)* 390
- Nicholson, B. 419
- Nikolajeva, M. 458
- Nishimura, Y. 50
- non-representational approach 190
- Nørgaard, N. 306
- normalisation 336
- Nottingham Healthcare Communication Corpus* 15
- Nottingham Multimodal Corpus (NMMC)* 21–22, 71, 73
- noun phrases 149
- nouns: gendered 387–390, 387; text and image sample analysis 97, 100–101, 102
- numerical identity 207
- NVivo* 54–55, 89
- object oriented thinking 206
- Observer XT* 75
- O'Connor, H. 51
- O'Dair, S. 506n1
- O'Donnell, M.B. 318
- O'Halloran, K.L. 85, 86
- O'Keefe, A. 150, 151
- Old Bailey Corpus* 112
- Oliver Twist* (Dickens) 503
- ontology 204, 205–208
- Open Persian Project* 411
- Operation War Diary* 533, 534
- optical character recognition (OCR) 419
- ordinary language philosophy 447
- Orlando Project, The* 479
- orthographic criterion 167
- orthographic types 165
- orthography 249
- O'Toole, M. 87–88, 88
- outer circle 542
- Outsourced Call Center Interactions corpus* 270
- Oxford English Corpus* 113
- Oxford English Dictionary (OED)* 117, 135, 174, 329, 556
- Oxford Learner's Dictionary of Academic English* 33
- Oxford Scholarly Editions Online* 406
- Oxford Text Archive (OTA)* 45
- Page, R. 188, 571
- Palfrey, J. 475
- Palmer, C.L. 478
- Palmer, H.E. 146, 149
- Pamphlet for Grammar* (Bullockar) 144
- Pandey, A.V. 477
- Pannapacker, W. 475
- Paradox of Choice, The* (RSAnimate) 467
- paralinguistic features 8
- parallel corpora **28**
- ParDeepBank* 408
- Parente, F., 503
- Paris Match* 87
- partial constructions 169
- Pasenek, B. 130
- passive auxiliaries 340, 341
- Paterson, L.L. 422
- Patrick, P. 284
- patterns: affirmative and negative **155, 156**; in corpus linguistics 389; in English literature research 495, 496; in narrative 318–319; in spoken corpora 20; in stylistic analysis 312; verb patterns 33, 42–43, **42**
- Paulus, T. 253, 254
- pauses 8
- Parsed Corpus of Early English Correspondence (PCEEC)* 332
- Pearson International Corpus of Academic English (PICAe)* 33
- Pechenick, E.A. 312
- pedagogy 323
- Pedersen, A. 50
- Pedersen, S. 51, 53
- Penn-parsed corpora 330
- Penn Phonetics Lab Forced Aligner (P2FA)* 75, 77
- Penn Treebank Project* 408, 409, 410
- Penn Treebank tagset 42
- Pennycook, A. 284

- Pentney, B. 51
 periodisation 334–336
 Perl 336
 perlocutionary effect 111
Perseid 413
Perseus collection of Greek and Roman materials 556
Perseus Digital Library 409–410, 411
 Pertejo, P.N. 331
 Petré, P. 333, 335, 340
 philosophy: background 440–443; chauvinism, worries about 450–452; conceptual analysis 441–442, 446; corpus research methods 449–450; critical issues 446–452; current contributions and research 443–445; digital humanities and 2; experimental 445–446, 449–450, 453n8, 453n10, 453n11; exploding intuitions 444; future directions 445–446; intuition-based enquiry 442–443; metaphilosophy 446; ordinary language philosophy 447; summary 452
 phrasal chaining 149
 phrasal verbs 148
 physics 444–445
Physics and Reality (Einstein) 502
 Piachaud, D. 421
 Picton, I. 323
 Piez, W. 468
 Pihlaja, S. 131, 139
 place-name co-occurrence (PNC) 422, 423, 423, 425, 426–427, 428, 430–431, **431**, 432, 433, 434, 435, 436
 point of view 462–465
 polyphony 466
 polysemy 171–172
 Pomerantz, A.M. 249
Portrait of a Lady, The (James) 503
 post-varieties approach 292
 Potter, J. 255
 Poutsma, H. 146
 poverty: conclusions 433–436; issues and research 420–421, **421**, 424, **425**, 429; research methods 422–424; sample analysis 424–433
Praat 9, 10, 76, 89, 250, 258n11, 290
Practical Study of Languages, The (Sweet) 146
 pragmaticalisation 118
 pragmatic annotation 10–11
 pragmatics: communicative 276–278; definition of 107; emergence of 108–109; perspective view 112; spoken corpora and 19–20; *see also* digital pragmatics.
Prague Czech Dependency Treebank 408
Prague Dependency Grammar 408
 predicates, secondary 343, 343
 Prenksy, M. 477
 Preparing DARIAH 560
Presentation of Self in Everyday Life, The (Goffman) 194
 presentism 447–448
 Presner, T. 212
 Preston, M. 29
 Prince, G. 462
 principle of total accountability 148
 Programming Historian 485
PROIEL project 407, 408
 projection structures 41–42, 43–44, **43**
 pronouns, personal 41, 42, 58–59, 60–61, 232–234, 546–551
 proper nouns 541
 proposals 42
 propositions 42, 448
 prosody: semantic 171, 182; speech 9, 68–69, 73, 77–79
ProsodyPro 75
Prosopography of Anglo-Saxon England (PAASE) 333
ProtAnt 234–235
 proto-idioms 169
 prototype approach 110
 Psathas, G. 244
 psycholinguistics 68, 494–495, 498–502, 506
 public humanities: background 531–534; communities of practice 535–536; corpus-based versus corpus-driven 539–540; corpus linguistics and 530–531, 538–541; critical issues and topics 534–538; crowdsourcing 534–535; current contributions and research 541–546; diachronic dimension 536–537; future directions 551–552; keywords 540–541; lexical features 544–546; lexical selection 537; methodological 537–538; research methods 538–541; sample analysis 546–551; Zooniverse *see* Zooniverse;
 punctuation 503–505
Python 224, 225, 336
 Quaglio, P. 18
 qualification 204
 qualitative methods 54
 quantification 203, 204
 Quantified Self Meetups 202, 204, 212
 Quantified Self Movement 205, 207, 208, 211
 quantitative perspectives 26, 309, 337–338
 quantitative-sociolinguistic work 284, 289
 QuickTime 250
 Quirk, R. 29, 147
 radical education movement 473
 Rampton, B. 198
 Ramsay, S. 484
Rbrul software 289–290
Reading Concordances (Sinclair) 311
 recontextualisation 214

- Red Riding Hood* 94
 reference corpus 170, 392, 540
 reflexivity 15
Reginald on Besetting Sins (Saki) 165
 regional corpora 10
 regression analysis 344, 345
 Reisigl, M. 222
 reliability of evidence 363–364
 repair 253
 repetition 313–318
Representative Corpus of English Historical Registers, A (ARCHER) 331
 reptile metaphors 135–138, 136, **137**, **138**
 Research Information Network Report 559
Research Unit for the Study of Variation, Contacts and Change in English (VARIENG) 31
 resemblance metaphors 137
 resemiotisation 208, 209, 214
 reverse inference 507n8
 Rheingold, H. 574
 rhetorical approaches 459
 Richard, M. 448
 Rieder, B. 126
 Robinson, K. 464
 Robinson, P. 139
 Rockwell, G. 309, 485
 Rogers, R. 483
 Röhle, T. 126
 Romanians 223–224, 228–232, **230**
 Romaniuk, T. 250
 Rooth, M. 73
 Rosenberg, A. 76
 Rosenbloom, P. 442
 Rosenbloom, P.S. 524
 Rosenzweig, R. 561, 563
 Ross, D. 444, 445, 446
Rossetti Archive 479
Routledge Handbook of Stylistics, The (Burke) 307
 RSAanimate 457, 460, 463–464, 465, 466, 467
R software 290
 Rühlemann, C. 14, 20, 151
 Rymes, B. 189

 Sacks, H. 242, 243, 244, 248
 Sag, I.A. 168
 Sairio, A. 32
 Saki 165
 Salahshour, N. 128
Salamanca Corpus 32
 Salami Project 559
 Salway, A. 89
 sample analyses: citizen science 546–551;
 corpus linguistics 390–399; cross-cultural communication 270–274;
 digital pragmatics 116–120; English language and history 424–433; English language and social media 575–581; English literature 503–505; ethnography 191–196; forensic linguistics 368–373; gesture studies 77–79; grammar 154–157; lexis 176–181; metaphor 133–138; multimodality 75–79; Mumsnet 57–62; poverty 424–433; public humanities 546–551; sociolinguistic 294–298; speech prosody 77–79; text and image 95–102
 sample approach 382–383
 SAMUELS tagger 175, 180
Santa Barbara Corpus of Spoken American English (SBCSAE) 7, 72, 73, 109
 Sayer, P. 276
 Sayers, J. 487
 Schauer, G.A. 117
 Schegloff, E.A. 243, 244, 248, 258n20
 Schembri, A. 381
 Schnobelen, T. 299
 Schöch, C. 288
 scholarly treebanks 410
 Schreibman, S. 107
 Schreier, D. 284
Scientific American 95–102
 Scollon, R. 276
 Scollon, S.B.K. 276
 ScopuS 511
 Scott, M. 32
 Scott, R. 503
 searchable talk 571
 search parameters 26–27
 Searle, J.R. 108, 111, 117
 secondary predicates 343, 343
 second language acquisition 153–154, **155**, **156**, 276–278
 Selber 468
 Selenium 224
 Selfe, C.L. 472
 self-harm 522–523, **522**
 selfies 191–196, 578
 self-quantification 203–215
 self-repair 253
 Selinker, L. 151
 semantic annotation 412
 semantic prosody 171, 182
 Semino, E. 312
 semiotisation 208, 209, 214
 Septuagint 407, 414n10
ServiceCORE 560
 sexism 51
Shakespeare Corpus 32
 Shakespeare's World 533
 Shakespeare, W. 506n3
 Shelley, E. 145
 Shohamy, E. 275, 277

- Short Introduction to English Grammar, with critical notes, A* (Lowth) 145
- Short, M. 307, 310, 312
- Sidnell, J. 248
- Siemens, R. 487
- signalling expressions 128
- Significance of Learners' Errors, The* (Corder) 151
- sign languages 381
- Silverstein, M. 191
- similarity-based paradigm 365
- Simpson, P. 307, 310, 322
- simulations 353
- sincerity conditions 111
- Sinclair, J. 29, 33, 147, 171, 311, 398, 537, 538, 539
- Sinclair, S. 309, 485
- Skelton, J.R. 269
- Sketch Engine* 7, 22, 27, 28, 45, 97, 226, 266, 539
- slang 178
- Slingerland, E. 453n12
- small correspondence corpora 39–44
- Small House at Allington, The* (Trollope) 321
- SmartKom* 73
- Smile or Die* (RSAnimate) 467
- Smith, B.A. 86
- Smith, R.C. 146
- Smith, S. 466
- Smithson, J. 51, 53
- soap operas 18
- social framework 246
- social media 73, 188, 520–523; *see also* English language and social media.
- Social Media and Everyday Politics* (Highfield) 196
- social pragmatics 107
- social semiotic approach 87, 88
- sociolinguistics: case study 294–298, **295**, 297; computational linguistics and 298–299; definition of 283–285; digital turn and 283, 285–298; forensic linguistics and 362; future directions 298–299; objects of study 285–288; online/offline relationship 292; research methods 289–291; research questions 291–294; spoken corpora and 13–14; variation and 285, 287, 288, 291–292; whiteboard animation and 459, 460–461
- Sociopragmatic Corpus, The* 117
- Software Environment for the Advancement of Scholarly Research (SEAR)* 559
- Sorlin, S. 307
- Sotirova, V. 307, 496
- Spanish 12
- Sparksight 457, 461
- spatial segregation analysis involves 433, 434, 435
- specialised corpora **28**
- speech acts 110–111, 113, 113–114, 115, 116–120
- speech-like varieties 17–18
- speech prosody: annotation 73; research on 68–69, 73; sample analysis 77–79; in spoken corpora 9
- Spencer-Oatey, H. 111
- Spilioti, T. 289
- Spinoza, B. 208
- Spitzer, L. 307
- Spoken British National Corpus (SBNC)* 6, 7, 13, 381, 385
- spoken corpora: academic analyses and 15–16; annotation of 10–11; applied linguistics and 15; critical issues and topics 12–21; data transcription and 8–10; English as lingua franca and 17; examples of 7; future directions 21–22; genres of 11–12; grammar and 8–9; healthcare communication and 15; learners and 16–17; monomodal 21; pedagogical aspect of 16–17; in pragmatics and discourse 19–20; representative 6–7; research methods 6–12; size of 7; sociolinguistics and 13–14; specialised 10; of speech-like varieties 17–18; speech prosody 68–69; teenage language and 14–15; variation across time 13; variations across national languages 12
- Spoken English Corpus (SEC)* 71
- spoken grammar 149
- Squires, L. 289
- Stamatatos, E. 365
- standard treebanks 410, 411
- Stanford Dependencies (SD)* 409
- Stanford Parser* 266
- Staples, S. 269
- statistical significance 238
- statistics 30
- Steele, C. 558, 559
- Stefanowitsch, A. 127
- Stenström, A.-B. 12, 109
- Stich, S. 453n19
- Stivers, T. 251
- Stockwell, P. 306, 307, 317
- Stokoe, E.H. 253, 254
- stylistic analysis: forensic 365–366; literary *see* literary stylistics;
- stylometry 308, 364–365, 495, 507n7
- suasive verbs 43
- subordination 149
- suicide prevention line 243
- Summey, D.C. 486
- surface reading 311
- Surman, M. 475
- surveillance capitalism 212
- Survey of English Usage (SEU)* 147

- Survey of Spoken English (SSE)* 147
 Svensson, P. 480, 511
 Swales, J. 15, 537
 Swan, M. 143
 Sweet, H. 146
 Swiss Short Message Service 575
Switchboard corpus 73
 symptomatic reading 311
 synchronicity 292
 synchronic variation 112
 syntax 143
 synthetic personalisation 465
 system 88
 systematicity 150
 systemic-functional grammar 87
 systemic functional linguistics (SFL) 571
Systems of Pragmatic Annotation in the Spoken Component of ICE-Ireland (SPICE-Ireland Corpus) 10
 Szmrecsanyi, B. 286, 299n3, 338

 Taavitsainen, I. 109, 115, 116
TaDiRAH resource 482
 Tagg, C. 211, 572
 taggers 41, 96, 96–97, 98, 175, 180, 330, 383
 Tagliamonte, S. 15, 291
 Tallant, J. 3, 444, 445, 446, 449
 Tanner, S. 419
 Tapscott, D. 458
 tautologies 119
 Taylor, C. 20
 Taylor, S. R. 193
 teaching materials, English 33–34
Teenage Health Freak corpus 521
 teenage language 14–15
 telephone call openings 150
 temporality 461–462
TenTen corpora 28
 Terras, M. 3, 86, 125, 144, 185, 215, 263, 310, 405, 450, 451, 456, 524
tertium comparationis 114
 Textal 309
 text and image: background 87–88; critical issues and topics 89–92; current contributions and research 88–89; future directions 103; hypothetical values **94**; limitations of digital tools 89–90; multimodal concordance **91**; multimodal corpora 90–92; research methods 92–95; sample analysis 95–102, **99**, **100**; visual content analysis 92–94; visual corpora 94–95
Text Creation Partnership (TCP) 175
 text editing 496–497, 506n2
Text Encoding Initiative (TEI) 40, 40, 173, 333
TextGrid 560
 text messages 292
 text types 147

 thanking 116–120
 ‘*The End of Theory*’ (Anderson) 213
 theoretical codes 56
 theoretical pragmatics 107
 theories of knowledge 209
The People’s Grammar (Shelley) 145
Thesaurus Linguae Graecae, The 413n5, 556
Thesaurus of Old English, A 135, 175
 Thewissen, J. 157
 Thibault, P.J. 90, 92
 thirdness/third place 277–278
 Thomas Aquinas 29
 Thomas, J. 269
 Thomas, M., 475
 Thompson, G. 29
 Thompson, S. 255
 Thurlow, C. 289
 Tibbo, H. 559
 Tiemensma, L. 458
 time alignment 9
 time-based data 89
 time series analysis 347–348, 349–350
Times, The 420–433, **421**, 423, 424, **425**, 428
TOEFL 2000 Spoken and Written Academic Language (T2k-SWAL) 33
 Tognini-Bonelli, E. 540
 tokens 165
 tone of voice 68–69
 Tono, Y. 152
 Toolan, M. 307, 308, 310
 topological phenomena 209
 Torgersen, E. 14
 Törnberg, A. 50
 Törnberg, P. 50
Toronto Teen Corpus: 2002–2006 (TTC) 15
 Tottie, G. 109
 traceability 110
Transana 250
 transcription: automated 256; conventions 257; conversation analysis and 248–249, 250–251, 254–255; grammar and 150–151; software applications for 250; of specialised corpora 10; of spoken corpora 8–9, 67
 translanguaging 573
 ‘*Translation and Translanguaging*’ 576
 trans-scripting 573
Treatise of Human Nature (Hume) 207
 treebanking 407, 409, 410
Trinity Lancaster Corpus (TLC) 16–17
 Trumper-Hecht, N. 277
 truth conditions 111
 tuning devices 128
 Turell, M.T. 368
 turn composition 245
 Turner, M. 125
 turn position 245
 turns construction 150

- Twitter* 66, 129, 131, 292, 295, **295**, 299n5, 571
 types 165
 typological phenomena 209
 typos 238
- UCREL Semantic Analysis System (USAS)* 175
 Underwood, T. 311
Universal Dependencies (UD) 409
 University of Birmingham 29
 University of Glasgow 563–564
 University of Nebraska-Lincoln 412, 413
 unmotivated looking 244, 250, 254, 258n6
 unscripted monologues 11
 Unsworth, J. 478, 559
 Ureña, J.M. 137
USAS tagger 383
- van Deursen, A. 488
 Van de Velde, F. 342
 van Dijck, J. 195, 196
 van Dijk, J.A.G.M. 488
 van Dijk, T. 220, 222
Van Gogh Letters Project 32
 Vanhouette, E. 186
 Van Leeuwen, T. 87, 88, 95
 van Leeuwen, T. 222
 van Ostade, Tieken-Boon 32
 van Peer, W. 494, 495, 496
VARBRUL 289
VARD 32
 variability-based neighbour clustering (VNC)
 350–351
 variationist approach 108
 variationist sociolinguistics 284, 285, 287
 variations: across national languages 12;
 across time 13; aspects of 337; diachronic
 112; linguistic 283–284, 292, 294–298;
 sociolinguistic 285, 287, 288, 291–292;
 speech-like 17–18; of spellings 172;
 synchronic 112
 variorum editing 496, 506n4
 Vee, A. 484
 Ventola, E. 89
 verb patterns 33, 42–43, **42**
 verbs: affirmative and negative patterns **155**,
 156; evidential 44; modal **155**, **156**, 157;
 phrasal 148; private 41; suasive 43; text and
 image sample analysis 100–101, 102
 vernacular spectacles 574
Vice Versa (Anstey) 318
Victorian Lives and Letters Consortium 31
VideoScribe 467
*Vienna Oxford Corpus of International
 English* 17
Vienna VOICE Corpus 17
 Vine, B. 267
 virtual visual synthetic personalisation 465
- VisMet* 130, 139
 visual/audio channels 288
 visual content analysis 92–94
 visual corpora 94–95
 visual ergonomics 507n10
 visual grammar 87
 visualisation: concordance as 311; digital
 literacies and 476–477; of *fog* in *Bleak House*
 317; literary stylistics and 310, 313, 322–323;
 mediated discourse analysis and 209–210,
 214; of whiteboard context 460–461
Visualizing English Print 32
 visual literacy 476–477
Visual Tagger 96–97, 97, 98
VLC 250
 vlogs 75–79
 voice 466–467, 517–523
*Volunteer and Crowdsourcing Economics
 Research (VOLCROWE)* 534–535
Voyant Tools 309, 315, 316
VU Amsterdam Metaphor Corpus 130
 Vygotsky, L.S. 203
- Waksman, S. 277
 Wales, K. 306, 307
 Waletsky, J. 459
 Wallace, D. 364
 Warwick, C. 559
 Watson, J. 466
Wearable Acoustic Metre (WAMs) 74
 wearable devices 74
*Web-as-a-Multimodal-Corpus Concordancing
 Tool* 75, 76, 76
 Webb, J. 421
WebCorp 382
 web corpora **28**
 weblogs 292
WebMAUS 290
 web scraping 224, 225
 Weedon, C. 54
 Weinberg, J. 453n19
 Weisser, M. 10
 Weller, M. 478, 487
 Wenger-Trayner, B. 535
 Wenger-Trayner, E. 535
weorðan 335, 340, 341
 Westlund, O. 221
 Whalen, M.R. 150
WhatsApp 575–581, 577, 578, 579, 580
 whiteboard animation 457; future directions
 467–468; gender inequity and 466; as
 narrative communication 456–457, 462;
 point of view 462–465; setting 460–461;
 temporality 461–462; voice 466–467
 White, D.S. 477
 Whiteley, S. 306, 307
 Wichmann, A. 9, 112

- Widdowson, H. 145
 Wiebesiek, L. Smakman, D. 284
Wikipedia 532
wildcards 299n6
 Williams, A.D. 458
 Willis, D. 33
 Wilson, A. 269
 Winner, L. 212
 Winters, J. 538
Wmatrix 71, 133, 139, 383
 Wodak, R. 222
 Wolfson, N. 114
 Wolk, C. 344, 345
Women Writers Online 479
 Wong, M.L.-Y. 117
 Woolls, D. 369
Woordenboek der Nederlandsche taal 329
WordNet 175
 words: definition of 167–169; distinguished
 from lexemes 165; frequency of 169–170; as
 proxies 166–167
WordSketch 34
WordSmith 20, 22, 90, 226, 266, 336, 539, 540
WordWanderer 317, 317
World Wide Web 73–74, 382
 Wright, D. 369, 370, 371
 written corpora: critical issues and topics 30;
 current contributions and research 31–34;
 diachronic/historical 31–33; English language
 teaching materials 33–34; future directions
 44–45; perspectives on 26; pre-loaded
 34–39; representative 30; sample analyses
 34–44; small correspondence 39–44; types of
 27–29, **28**
 Wulf, W.A. 558

XML 39–40, 67, 90, 175, 309

 Yano, C. 50
*York-Helsinki Parsed Corpus of Old English
 Poetry* 32
*York-parsed corpus of Old English prose
 (YCOE)* 330
York Poetry corpus (YPC) 330
YouTube 66, 76, 131, 139, 189

 Zappavigna, M. 571, 578
 Zhao, S. 578
 Zimmerman, D.H. 150
 Zipfian curve 169, 170
Zooniverse: Ancient Lives 546–551; as citizen
 science model 533–534; communities of
 practice 535–536; corpus linguistics and 539;
 crowdsourcing 534–535; crowdsourcing
 and 551; ‘Dark Matter’ corpus 544–546;
 definition of 530; diachronic dimension
 536–537; English as *lingua franca* 541–543;
 lexical selection 537; overview 532–533;
 roles within 536
 Zundert, J. van 484
Zurich English Newspaper Corpus (ZEN) 31
 Zyngier, S. 307, 323



Taylor & Francis Group
an informa business



Taylor & Francis eBooks

www.taylorfrancis.com

A single destination for eBooks from Taylor & Francis with increased functionality and an improved user experience to meet the needs of our customers.

90,000+ eBooks of award-winning academic content in Humanities, Social Science, Science, Technology, Engineering, and Medical written by a global network of editors and authors.

TAYLOR & FRANCIS EBOOKS OFFERS:

A streamlined experience for our library customers

A single point of discovery for all of our eBook content

Improved search and discovery of content at both book and chapter level

REQUEST A FREE TRIAL
support@taylorfrancis.com

 **Routledge**
Taylor & Francis Group

 **CRC Press**
Taylor & Francis Group