The Routledge Handbook of Corpus Linguistics

The Routledge Handbook of Corpus Linguistics 2e provides an updated overview of a dynamic and rapidly growing area with a widely applied methodology. Over a decade on from the first edition of the Handbook, this collection of 47 chapters from experts in key areas offers a comprehensive introduction to both the development and use of corpora as well as their ever-evolving applications to other areas, such as digital humanities, sociolinguistics, stylistics, translation studies, materials design, language teaching and teacher development, media discourse, discourse analysis, forensic linguistics, second language acquisition and testing.

The new edition updates all core chapters and includes new chapters on corpus linguistics and statistics, digital humanities, translation, phonetics and phonology, second language acquisition, social media and theoretical perspectives. Chapters provide annotated further reading lists and step-by-step guides as well as detailed overviews across a wide range of themes. The Handbook also includes a wealth of case studies that draw on some of the many new corpora and corpus tools that have emerged in the last decade.

Organised across four themes, moving from the basic start-up topics such as corpus building and design to analysis, application and reflection, this second edition remains a crucial point of reference for advanced undergraduates, postgraduates and scholars in applied linguistics.

Anne O’Keeffe is senior lecturer at MIC, University of Limerick, Ireland. Her publications include the titles From Corpus to Classroom (2007), English Grammar Today (2011), Introducing Pragmatics in Use (2nd edition 2020) and, as co-editor, The Routledge Handbook of Corpus Linguistics (1st edition 2010). With Geraldine Mark, she was co-Principal Investigator of the English Grammar Profile. She is co-editor, with Michael J. McCarthy, of two book series: The Routledge Corpus Linguistics Guides and The Routledge Applied Corpus Linguistics.

Michael J. McCarthy is emeritus professor of applied linguistics, University of Nottingham. He is (co)author/(co)editor of 57 books, including Touchstone, Viewpoint, The Cambridge Grammar of English, English Grammar Today, From Corpus to Classroom, Innovations and Challenge in Grammar and titles in the English Vocabulary in Use series. He is author/co-author of 120 academic papers. He was co-founder of the CANCODE and CANBEC spoken English corpora projects. His recent research has focused on spoken grammar. He has taught in the UK, Europe and Asia and has been involved in language teaching and applied linguistics for 56 years.
Routledge Handbooks in Applied Linguistics

*Routledge Handbooks in Applied Linguistics* provide comprehensive overviews of the key topics in applied linguistics. All entries for the handbooks are specially commissioned and written by leading scholars in the field. Clear, accessible and carefully edited *Routledge Handbooks in Applied Linguistics* are the ideal resource for both advanced undergraduates and postgraduate students.

The Routledge Handbook of Corpus Approaches to Discourse Analysis
*Edited by Eric Friginal and Jack A. Hardy*

The Routledge Handbook of World Englishes
Second Edition
*Edited by Andy Kirkpatrick*

The Routledge Handbook of Language, Gender and Sexuality
*Edited by Jo Angouri and Judith Baxter*

The Routledge Handbook of Plurilingual Language Education
*Edited by Enrica Piccardo, Aline Germain-Rutherford and Geoff Lawrence*

The Routledge Handbook of the Psychology of Language Learning and Teaching
*Edited by Tammy Gregersen and Sarah Mercer*

The Routledge Handbook of Language Testing
Second Edition
*Edited by Glenn Fulcher and Luke Harding*

The Routledge Handbook of Corpus Linguistics
Second Edition
*Edited by Anne O’Keeffe and Michael J. McCarthy*

For a full list of titles in this series, please visit www.routledge.com/series/RHAL
For Ron Carter, whose insight, humour and friendship we will forever miss.
Contents

List of illustrations xii
List of contributors xvi
Acknowledgements xxviii

1 ‘Of what is past, or passing, or to come’: corpus linguistics, changes and challenges 1
   Anne O’Keeffe and Michael J. McCarthy

PART I
Building and designing a corpus: the basics 11

2 Building a corpus: what are key considerations? 13
   Randi Reppen

3 Building a spoken corpus: what are the basics? 21
   Dawn Knight and Svenja Adolphs

4 Building a written corpus: what are the basics? 35
   Tony McEnery and Gavin Brookes

5 Building small specialised corpora 48
   Almut Koester

6 Building a corpus to represent a variety of a language 62
   Brian Clancy

7 Building a specialised audiovisual corpus 75
   Paul Thompson

8 What corpora are available? 89
   Martin Weisser
9 What can corpus software do? 103
   Laurence Anthony

10 What are the basics of analysing a corpus? 126
   Christian Jones

11 How can a corpus be used to explore patterns? 140
   Susan Hunston

12 What can corpus software reveal about language development? 155
   Xiaofei Lu

13 How to use statistics in quantitative corpus analysis 168
   Stefan Th. Gries

PART II
Using a corpus to investigate language 183

14 What can a corpus tell us about lexis? 185
   David Oakey

15 What can a corpus tell us about multi-word units? 204
   Chris Greaves and Martin Warren

16 What can a corpus tell us about grammar? 221
   Susan Conrad

17 What can a corpus tell us about registers and genres? 235
   Bethany Gray

18 What can a corpus tell us about discourse? 250
   Gerlinde Mautner

19 What can a corpus tell us about pragmatics? 263
   Christoph Rühlemann

20 What can a corpus tell us about phonetic and phonological variation? 281
   Alexandra Vella and Sarah Grech
PART III
Corpora, language pedagogy and language acquisition 297

21 What can a corpus tell us about language teaching? 299
   Winnie Cheng and Phoenix Lam

22 What can corpora tell us about language learning? 313
   Pascual Pérez-Paredes and Geraldine Mark

23 What can corpus linguistics tell us about second language acquisition? 328
   Ute Römer and Jamie Garner

24 What can a corpus tell us about vocabulary teaching materials? 341
   Martha Jones and Philip Durrant

25 What can a corpus tell us about grammar teaching materials? 358
   Graham Burton

26 Corpus-informed course design 371
   Jeanne McCarten

27 Using corpora to write dictionaries 387
   Geraint Rees

28 What can corpora tell us about English for Academic Purposes? 405
   Oliver Ballance and Averil Coxhead

29 What is data-driven learning? 416
   Angela Chambers

30 Using data-driven learning in language teaching 430
   Gaëtanelle Gilquin and Sylviane Granger

31 Using corpora for writing instruction 443
   Lynne Flowerdew

32 How can corpora be used in teacher education? 456
   Fiona Farr

33 How can teachers use a corpus for their own research? 469
   Elaine Vaughan
PART IV
Corpora and applied research

34 How to use corpora for translation
   Silvia Bernardini

35 Using corpus linguistics to explore the language of poetry: a
   stylometric approach to Yeats’ poems
   Dan McIntyre and Brian Walker

36 Using corpus linguistics to explore literary speech representation:
   non-standard language in fiction
   Carolina P. Amador-Moreno and Ana Maria Terrazas-Calero

37 Exploring narrative fiction: corpora and digital humanities
   projects
   Michaela Mahlberg and Viola Wiegand

38 Corpora and the language of films: exploring dialogue in English
   and Italian
   Maria Pavesi

39 How to use corpus linguistics in sociolinguistics: a case study of
   modal verb use, age and change over time
   Paul Baker and Frazer Heritage

40 Corpus linguistics in the study of news media
   Anna Marchi

41 How to use corpus linguistics in forensic linguistics
   Mathew Gillings

42 Corpus linguistics in the study of political discourse: recent
   directions
   Charlotte Taylor

43 Corpus linguistics and health communication: using corpora to
   examine the representation of health and illness
   Gavin Brookes, Sarah Atkins and Kevin Harvey

44 Corpus linguistics and intercultural communication: avoiding the
   essentialist trap
   Michael Handford
45  Corpora in language testing: developments, challenges and opportunities
    Sara T. Cushing  

46  Corpus linguistics and the study of social media: a case study using multi-dimensional analysis
    Tony Berber Sardinha  

47  Posthumanism and corpus linguistics
    Kieran O’Halloran  

Index  

xii
Illustrations

Figures

3.1 An example of transcribed speech, taken from the NMMC
3.2 A column-based transcript
3.3 Time-based tracks in ELAN
7.1 The ELAN programme interface (showing a sample from the ACLEW project, https://psyarxiv.com/bf63y/): video, annotations with timestamps and waveforms
7.2 Screenshot of Glossa interface, showing results of a search on the Norwegian Big Brother corpus
9.1 Screenshot of AntConc displaying frequency values for all the words in the ICNALE corpus
9.2 Screenshot of AntConc showing a standard frequency-based keyword list for the Japanese component of the ICNALE written corpus
9.3 (a and b) Screenshots of AntConc displaying clusters in the ICNALE corpus
9.4 (a and b) Screenshots of AntConc displaying n-grams in the ICNALE corpus
9.5 Screenshot of AntConc displaying the collocates of smoking in the ICNALE corpus
9.6 Screenshot of AntConc displaying the concordance for the search term smoking in the ICNALE corpus
9.7 (a and b) Screenshots of AntConc displaying concordance plots for the phrase banning smoking in the ICNALE corpus
9.8 (a and b) TagAnt POS tagging texts and AntConc displaying a wordlist for the ICNALE corpus tagged by part of speech
9.9 Screenshot of ProtAnt after ranking texts of the ICNALE corpus by their use of academic words
13.1 Dendrogram of the data discussed in Divjak and Gries (2006)
14.1 Different ways of graphically presenting a word frequency list for The Adventures of Sherlock Holmes
14.2 Concordance lines for remote in COCA
14.3 Sample concordance lines for want in COCA sorted one, two and three words to the right of the search word
14.4 Adjectives modified by severely in COCA and English Web 2020 on Sketch Engine
14.5 Verbs for which *diversity* is the object in COCA and English Web 2020 on Sketch Engine
14.6 Concordance lines for *crescendo* as the object of *reach* in COCA
15.1 Sample concordance lines for “expenditure/reduce”
15.2 Sample concordance lines for the collocational framework *the..... of the*
15.3 Sample concordance lines for the meaning shift unit “play/role”
15.4 Sample concordance lines for the organisational framework *I think..... because*
16.1 Conditions associated with omission of *that* in *that*-clauses
17.1 Distribution of RA types in six disciplines along Dimension 2 “Contextualized Narrative Description” vs. “Procedural Description” (adapted from Gray 2015)
19.1 Number of distinct word types in 7- to 12-word turns in BNC-C
19.2 Distribution of inserts, function words and content words in a sample of 1,000 turns from 7- to 12-word turns in the BNC-C
19.3 Finger-bunch gesture depicting the word “location”
20.1 The composition of the spoken component of the BNC
22.1 Developmental profiling techniques used in Vyatkina (2013)
24.1 Configurations list of two-word concgrams
24.2 Phrases including *flow and rate*
24.3 Phrases including *rate and flow*
26.1 Collocates of *see* in BWAWE and *Spoken BNC* (2014)
26.2 Frequency of days of the week in the *Spoken BNC* (2014)
27.1 Concordance lines for “perpetrate” in the BNC 1994 shown in Sketch Engine
27.2 Approximate text proportions in the BNC 1994
27.3 Partial word sketch for the verb “fire” in the BNC 1994 using Sketch Engine
27.4 Word sketch difference for “fire” in the BNC 1994 using Sketch Engine
33.1 Some contexts of professional interaction
35.1 Cluster analysis of Yeats’ poetry in 13 volumes based on $n$MFW = 100
35.2 Cluster analysis of Yeats’ poetry in 13 volumes based on $n$MFW = 1,000
35.3 PCA of 13 volumes of Yeats’ poetry $n$MFW = 100
35.4 PCA of 13 volumes of Yeats’ poetry $n$MFW = 1000
35.5 PCA of 13 volumes of Yeats’ poetry $n$MFW = 200 with loadings
35.6 Bootstrap consensus tree of Yeats’ poetry in 13 volumes based on $n$MFW = 100 to 1,000 in increments of 50
36.1 Colligation patterns across the books per 100 tokens
36.2 Raw token count illustrating prosodic representation of NISo
37.1 10 Concordances lines of *Daisy* in *David Copperfield* (retrieved with CLiC)
37.2 Distribution plots for *Daisy, Doady and Davy* in *David Copperfield* (retrieved with CLiC)
37.3 The top 20 -ing forms in DNov long suspensions (retrieved with CLiC)
39.1 Diachronic change in *may* as a modal verb across comparable age groups
39.2 Diachronic changes in how frequently decade of birth cohorts used the modal verb *may*
568

39.3 Percentages of the functions of *may* in three groups
570

46.1 Means for platform dim. 1 ($R^2 = 0.2\%$; $F = 36.4$; $p < 0.0001$)
665

46.2 Means for user group dim. 1 ($R^2 = 13.04\%;$ $F = 3198.53; p < 0.0001$)
665

46.3 Means for user dim. 1 ($R^2 = 26.4\%;$ $F = 70.03; p < 0.0001$)
666

46.4 Means for platform dim. 2 ($R^2 = 0.98\%;$ $F = 211.84; p < 0.0001$)
669

46.5 Means for user group dim. 2 ($R^2 = 2.35\%;$ $F = 514.873; p < 0.0001$)
669

46.6 Means for user dim. 2 ($R^2 = 14.36\%;$ $F = 32.8; p < 0.0001$)
670

47.1 Augmented reality for analysing data with IBM immersive insights
677

47.2 Diffraction patterns
678

47.3 Key semantic domains for chapter 18 *Ulysses* – generated by *WMatrix*
682

47.4 Screenshot of FDA text on FDA website
686

47.5 Dominant conceptual framings across FDA text
687

47.6 Key semantic domains for PETA “animal testing” corpus
688

Tables

5.1 Total number of occurrences of modals of obligation in each macro-genre (per thousand word)
57

8.1 The composition of the *Brown Corpus*
94

10.1 The frequency of *I mean* per million words in informal TV shows across time using the TV corpus
128

10.2 Sample frequency lists
129

10.3 Positive keywords in USTC (B2) compared with LINDSEI as a reference corpus
131

10.4 Four-word n-grams in the CLiC Dickens corpus
133

10.5 Four-word n-grams from USTC at the B2 level
135

10.6 The priming of divorced in academic and fictional texts
137

11.1 Rewriting with the pattern v-link ADJ about n
150

13.1 A co-occurrence table based on the frequencies of co-occurrence (and assuming a corpus size of 100,000 constructions, however defined)
172

14.1 Excerpts from word frequency lists for *The Adventures of Sherlock Holmes* in *Sketch Engine*
188

14.2 Sample from the word frequency list for the *Corpus of Contemporary American English* (COCA)
190

14.3 Excerpt from the word sketch for *remote* based on the English Web 2015 corpus in *Sketch Engine*
195

20.1 General corpora and their spoken component
285

20.2 Corpora with a spoken component and transcripts
286

20.3 Corpora with a spoken component, transcripts and audio (in some cases time-aligned)
287

22.1 Corpus design elements and considerations
316

22.2 A sample of learner corpora and their research purposes and content
317

22.3 Development of past simple form and functions across proficiency levels
321

22.4 Extracts from the *English Grammar Profile* of past simple development
321
24.1 List of the most frequent keywords in the science and engineering corpus according to log likelihood (LL) values 349
24.2 List of 50 keywords selected from the science and engineering corpus considered to be useful or very useful 350
26.1 Examples of chunks as expressions with unitary meaning from the Spoken BNC (2014) 379
31.1 Concordance lines for “delimiting the case under consideration” (adapted from Weber 2001: 17) 447
34.1 The top 20 collocates of contamination and contaminazione in two web corpora of French and Italian [with English glosses in square brackets] 491
34.2 Statistical significance and effect size values for the comparison of the frequencies of the two patterns in original and translated English 493
35.1 The Yeats Corpus of Poetry 505
35.2 The Early Yeats Corpus: volumes included and totals 509
35.3 The Late Yeats Corpus: volumes included and totals 510
35.4 Keywords in the Early Yeats Corpus 510
35.5 Keywords in the Late Yeats Corpus 511
36.1 The RO’CK corpus texts 522
36.2 Raw token count per book 523
36.3 Raw distribution of tokens by gender and book 527
37.1 Top 20 forms ending with ing in DNov long suspensions (bold – in top 20 of all three corpora; italics – only in top 20 of that corpus; * – keywords in DNov long suspensions vs. 19C and ChiLit combined) 541
38.1 Bilingual concordances for a four-word cluster from the PCFD (Freddi 2011: 152–5; back translations in brackets) 553
38.2 Bilingual concordances for second person pronouns from the PCFD 553
38.3 Bilingual concordances for Italian adversative ma “but” from the PCFD 557
38.4 Concordances for mate and man from the PCFD 558
39.1 Age cohorts and how old members of those cohorts would have been in the BNC1994 and BNC2014 566
39.2 Common phraseological units containing may for different age groups 571
43.1 Sexual health keywords in AHEC, ranked by log-likelihood score 619
43.2 VIOLENCE metaphorical collocates of dementia (MI ≥ 3), ranked by collocation frequency (brackets) 623
44.1 Example CLIC steps 636
44.2 Checklist for conducting CLIC 638
44.3 JAICEE “different cultural” concordances 639
46.1 Corpus used in the study: breakdown by platform 661
46.2 Corpus used in the study: breakdown by user group 661
46.3 Interval and ordinal scale example 662
46.4 Factor pattern for dimension 1 664
46.5 Factor pattern for dimension 2 668
47.1 Frequencies for words under the key semantic domain SMOKING_and_NON-MEDICAL_DRUGS 688
Contributors

Svenja Adolphs is a professor of English language and linguistics at the University of Nottingham, UK. Her research interests are in multimodal spoken corpus linguistics, corpus-based pragmatics and discourse analysis. Along with Dawn Knight, Adolphs recently edited The Routledge Handbook of English Language and Digital Humanities (2020).

Carolina P. Amador-Moreno is a professor of English linguistics at the University of Bergen. Her research interests centre on the English spoken in Ireland and include stylistics, discourse analysis, corpus linguistics, sociolinguistics and pragmatics. She is the author, among others, of Orality in Written Texts: Using Historical Corpora to Investigate Irish English (1700–1900), Routledge (2019); An Introduction to Irish English, Equinox (2010); the co-edited volumes Irish Identities: Sociolinguistic Perspectives, Mouton de Gruyter (2020); Voice and Discourse in the Irish Context, Palgrave-Macmillan (2017); Pragmatic Markers in Irish English (2015), John Benjamins; and Fictionalising Orality, a special issue of the journal Sociolinguistic Studies (2011).

Laurence Anthony is a professor of applied linguistics at the Faculty of Science and Engineering, Waseda University, Japan, where he is also the Director of the Center for English Language Education (CELESE). He has a BSc degree (mathematical physics) from the University of Manchester, UK, and MA (TESL/TEFL) and PhD (applied linguistics) degrees from the University of Birmingham, UK. His main research interests are in corpus linguistics, educational technology and English for specific purposes (ESP). He received the National Prize of the Japan Association for English Corpus Studies (JAECs) in 2012 for his work in corpus software design.

Sarah Atkins is a research fellow at the Aston Institute for Forensic Linguistics, Aston University, where she leads on a number of external projects and partnerships. She also holds a visiting research fellowship at the Centre for Sustainable Working Life, Birkbeck, University of London. She has conducted research in a range of professional settings, most notably in health care, with an emphasis on applying findings to practice. Her work on communication skills training in medical education, which combines corpus linguistic and micro-analytic approaches to spoken interaction, has resulted in a range of policy applications and award-winning workshops for health care professionals.

Paul Baker is a professor of English language at Lancaster University. He has written 20 books on various aspects of language, identity, discourse and corpus linguistics.
These include *Sociolinguistics and Corpus Linguistics* and *Using Corpora to Analyse Gender*. He is commissioning editor of the journal *Corpora* and a fellow of the Royal Society of Arts. His latest book, *Fabulosa: The Story of Polari, Britain’s Secret Gay Language*, was a *Times Literary Supplement* Book of the Year in 2019.

**Oliver Ballance** is a lecture in Applied Linguistics and English for Academic Purposes in the School of Humanities, Media and Creative Communication at Massey University, New Zealand. Oliver teaches courses on EAP and curriculum design and supervises postgraduate research projects. His own research interests are focused upon the interface between corpus linguistics and language for specific purposes.

**Silvia Bernardini** is a professor of English linguistics at the Department of Interpreting and Translation of the University of Bologna, Italy, where she teaches translation from English into Italian and corpus linguistics. She has published widely on corpus use in translator education and for translation practice and research. Her research interests include the investigation of the points of contact between translation and interpreting and translation and non-native writing, seen as instances of bilingual language use.

**Gavin Brookes** is research fellow in the Department of Linguistics and English Language at Lancaster University. He is Associate Editor of the *International Journal of Corpus Linguistics* (John Benjamins) and Co-Editor of the *Corpus and Discourse* book series (Bloomsbury). Gavin has published widely on corpus linguistics, (critical) discourse studies and health communication, and is the author of *Corpus, Discourse and Mental Health* (with Daniel Hunt, Bloomsbury, 2020) and *Obesity in the News: Language and Representation in the Press* (with Paul Baker, Routledge, 2021).

**Graham Burton** is a researcher and lecturer at the Faculty of Education, Free University of Bozen-Bolzano and has written coursebooks and other teaching materials for a number of publishers. His PhD focused on how the consensus on pedagogical grammar for ELT evolved, how it is sustained and how it compares to empirical data on learner language.

**Angela Chambers** is a professor emerita of applied languages at the University of Limerick. She has taught in universities in France, the United Kingdom and Ireland. Her research interests focus on the use of corpora in language learning. She has published extensively in this area, including articles in journals such as *Language Learning & Technology, ReCALL, the International Journal of Corpus Linguistics, the Revue Française de Linguistique Appliquée* and *Language Teaching*. In 2007 she was guest editor for a special issue of *ReCALL* on corpora in language learning. She currently teaches academic writing at the postgraduate and postdoctoral level.

**Winnie Cheng** is former professor and former Director of the Research Centre for Professional Communication in English (RCPCE) of the Department of English at The Hong Kong Polytechnic University. Her research interests include corpus linguistics, conversation analysis, critical discourse analysis, discourse intonation, EAP, ESP, intercultural communication, pragmatics and writing across the curriculum.
**Brian Clancy** is currently a lecturer in applied linguistics at Mary Immaculate College, Ireland. His research work focusses on the blend of a corpus linguistic methodology with the discourse analytic approaches of pragmatics and sociolinguistics. His primary methodological interests relate to the use of corpora in the study of language varieties and the construction and analysis of small corpora. His published work explores language use in intimate settings, such as between family and close friends, and the language variety Irish English. He is the author of *Investigating Intimate Discourse: Exploring the Spoken Interaction of Families, Couples and Close Friends* (Routledge, 2016) and co-authored *Introducing Pragmatics in Use* (Routledge, 2011 and 2020).

**Susan Conrad** is a professor of applied linguistics at Portland State University, Portland, Oregon, USA. She has used corpus linguistics to study English grammar in a variety of contexts, from general conversation to engineering. Her publications include the *Grammar of Spoken and Written English, The Cambridge Introduction to Applied Linguistics, Real Grammar* and *Register, Genre, and Style*. She coordinates the Civil Engineering Writing Project, in which corpus linguists and engineers collaborate to improve students' writing skills. Her teaching experiences in southern Africa, South Korea and the United States convinced her corpus techniques were useful long before they were well known.

**Averil Coxhead** is a professor at Te Herenga Waka/Victoria University of Wellington, Aotearoa/New Zealand where she teaches undergraduate and postgraduate courses in applied linguistics and TESOL. Averil is the author of *Vocabulary and English for Specific Purposes Research* (Routledge, 2018) and co-author of *English for Vocational Purposes* (Routledge, 2020) and a textbook series with Professor Paul Nation entitled *Reading for the Academic World* (Seed Learning, 2018). Her current research includes technical multiword units in trades education, vocabulary in hip hop (with Friederike Tegge) and wordlists in trades such as carpentry and plumbing in English and their translation into Tongan with Falakiko Tu’amoheleoa.

**Sara T. Cushing** is a professor of applied linguistics at Georgia State University. She received her PhD in applied linguistics from UCLA. She has published research in the areas of assessment, second language writing and teacher education. She has been invited to speak and conduct workshops on second language writing assessment throughout the world, most recently in Vietnam, Colombia, Thailand and Norway. Her current research focuses on assessing integrated skills, the use of automated scoring for second language writing and applications of corpus linguistics to assessment.

**Philip Durrant** is an associate professor in language education at the University of Exeter. He has been a language teacher and researcher for over 20 years, working at schools and universities in the UK and Turkey. He has published widely on corpus linguistics, vocabulary learning and academic writing.

**Fiona Farr** is an associate professor of applied linguistics and TESOL at the University of Limerick. She is the Director of CALS (Centre for Applied Language Studies) and Research Director in MLAL. Her research interests include teacher education and professional development, applied corpus linguistics; and language learning and...
technology. She is the author of *Teaching Practice Feedback: An Investigation of Spoken and Written Modes* (2011), *Practice in TESOL* (2015) and *Social Interaction in Language Teacher Education* (2019, with Farrell and Riordan). She is the co-editor of the *EUP Textbooks in TESOL* series and of the *Routledge Handbook of Language Learning and Technology* (2016).

**Lynne Flowerdew** is currently a visiting research fellow in the Department of Applied Linguistics and Communication, Birkbeck, University of London. Her main research and teaching interests include corpus linguistics, discourse analysis, EAP/ESP and disciplinary writing. She has published widely in these areas in international journals and prestigious edited collections and has also authored and co-edited several books.

**Jamie Garner** is a lecturer in the Department of Linguistics at the University of Florida, where she also serves as the coordinator of the undergraduate Linguistics and undergraduate Teaching English as a Second Language (TESL) programs. Her current research interests include learner corpus research, phraseology and second language acquisition. Her research has been published in the *International Journal of Learner Corpus Research, System, The Modern Language Journal, System* and *International Review of Applied Linguistics in Language Teaching*.

**Mathew Gillings** is an assistant professor at the Vienna University of Economics and Business. He completed his PhD at the ESRC Centre for Corpus Approaches to Social Science at Lancaster University, where he explored verbal cues to deception through the use of corpus-based methods. His research interests are in various aspects of corpus linguistics, but most recently he has applied the method to the study of deception, politeness and Shakespeare’s language.

**Gaëtanelle Gilquin** is a professor of English language and linguistics at the University of Louvain. She is the coordinator of LINDSEI (Louvain International Database of Spoken English Interlanguage) and PROCEED (PROcess Corpus of English in EDucation) and one of the editors of the *Cambridge Handbook of Learner Corpus Research*. Her research interests include the use of native and learner corpora for the description and teaching of language, the analysis of the writing process through screencasting and keylogging and the comparison of learner Englishes and world Englishes.

**Sylviane Granger** is an emerita professor of English language and linguistics at the University of Louvain. In 1990 she launched the first large-scale learner corpus project, the *International Corpus of Learner English*, and since then has played a key role in defining the different facets of the field of learner corpus research. Her current research interests focus on the analysis of phraseology in native and learner language and its integration into reference and instructional materials. One of her latest publications is *Perspectives on the L2 Phrasicon: The View from Learner Corpora* (2021).

**Bethany Gray** is an associate professor of English (Applied Linguistics and Technology program) at Iowa State University. Her research investigates phraseological, grammatical and lexico-grammatical variation across registers of English, with a particular focus on disciplinary variation in academic writing and the development of grammatical complexity in novice and L2 writers. She is a co-founding editor of *Register Studies*.
(John Benjamins), and her work appears in journals such as *Applied Linguistics*, *TESOL Quarterly*, *Journal of English for Academic Purposes*, *Corpora* and *International Journal of Corpus Linguistics*. Her books include *Linguistic Variation in Research Articles* (John Benjamins) and *Grammatical Complexity in Academic English* (Cambridge University Press, co-authored with Douglas Biber).

**Chris Greaves** was a senior research fellow in the English Department of the Hong Kong Polytechnic University until his retirement in 2012. He is well-known for his groundbreaking corpus linguistics software such as ConcApp, iConc and ConcGram and his research in the areas of corpus linguistics and phraseological variation. Sadly, Chris passed away in December 2020 before this volume went into production.

**Sarah Grech** holds a PhD in linguistics from the University of Malta. She is a senior lecturer with the Institute of Linguistics and Language Technology, University of Malta. Her research interests are in sociophonetics and language variation and change. She is currently leading or co-coordinating projects on Maltese English, as well as learner English in Malta, in order to identify key features of this variety of English within Malta's multilingual context.

**Stefan Th. Gries** is a professor of linguistics in the Department of Linguistics at the University of California, Santa Barbara (UCSB) and Chair of English Linguistics (corpus linguistics with a focus on quantitative methods, 25 per cent) at the Justus-Liebig-Universität Giessen. He was a Visiting Chair (2013–17) of the Centre for Corpus Approaches to Social Science at Lancaster University, visiting professor at five LSA Linguistic Institutes between 2007 and 2019 and the Leibniz Professor (spring semester 2017) at the Research Academy Leipzig of the University of Leipzig.

**Michael Handford** is a professor of applied linguistics at Cardiff University, where he is Director of internationalisation for the School of English, Communication and Philosophy. His research interests include discourse in professional settings, cultural identities at work, the application of corpus tools in discourse analysis, using corpora for the analysis of intercultural communication, essentialism and stereotyping, engineering education and communication, diversity and education and internationalisation in higher education. He is the author of *The Language of Business Meetings*, published by Cambridge University Press, and co-editor with J.P. Gee of the *Routledge Handbook of Discourse Analysis*.

**Kevin Harvey** is an associate professor in sociolinguistics in the School of English at the University of Nottingham. His principal research interests lie in the field of applied discourse analysis, multi-modal discourse analysis and corpus linguistics. Broadly, he is interested in interdisciplinary approaches to professional communication, with a special emphasis on health communication. He is editor of the dementia blog, *Dementia Day-to-Day*, and currently runs a series of dementia reading groups in care homes and local hospitals. His books include *Investigating Adolescent Health Communication: A Corpus Linguistics Approach* (Bloomsbury, 2013) and *Exploring Health Communication: Language in Action* (with Nelya Koteyko, Routledge, 2013).
Frazer Heritage holds a PhD in linguistics from Lancaster University and is currently an assistant lecturer in the Department of Psychology at Birmingham City University. His research focuses on corpus approaches to both critical discourse analysis and critical sociolinguistics. Frazer is interested in the linguistic construction and representation of gender, sexuality, and identity in different corpora, particularly in corpora of videogames and of language from social media. He is the author of the monograph Language, Gender and Videogames: Using Corpora to Analyse the Representation of Gender in Fantasy Videogames (Palgrave Macmillan, 2021).

Susan Hunston is a professor of English language at the University of Birmingham, where she has worked since 1996. Her research focuses on corpus linguistics, especially the lexis–grammar interface, and on discourse analysis, especially evaluative language and academic discourse. She has written several books, most recently Interdisciplinary Research Discourse (2019, Routledge) with Paul Thompson, and numerous articles. Susan has held lectureships at Mindanao State University, the National University of Singapore and the University of Surrey and has also worked as a grammarian on the COBUILD project.

Christian Jones is a senior lecturer in TESOL and applied linguistics at the University of Liverpool. His main research interests are connected to spoken language, and he has published research related to spoken corpora, lexis, lexico-grammar and instructed second language acquisition. He is the co-author (with Daniel Waller) of Corpus Linguistics for Grammar: A Guide for Research (Routledge, 2015), Successful Spoken English: Findings from Learner Corpora (with Shelley Byrne and Nicola Halenko) (Routledge, 2017) and editor of Literature, Spoken Language and Speaking Skills in Second Language Learning, a collection of research on this area (Cambridge University Press, 2019).

Martha Jones is an assistant professor in the School of Education at the University of Nottingham. She is the MA Teaching English for Academic Purposes (web-based) course leader and teaches on the face-to-face and web-based MA Teaching English to Speakers of Other Languages programme. Martha's publications have focused on the development and exploitation of corpora for teaching and research purposes, the acquisition of academic and disciplinary vocabulary, the use of ePortfolios in teacher education and the development of e-resources, including an open educational resource. She has also conducted research using combined corpus and genre-based approaches focusing on disciplinary written discourse.

Dawn Knight, Cardiff University, has expertise in corpus linguistics, discourse analysis, digital interaction and non-verbal communication. She was Principal Investigator on the £1.8m ESRC/AHRC-funded National Corpus of Contemporary Welsh (CorCenCC) and Welsh government–funded Cross-lingual Word-Embeddings projects. Knight has experience constructing and utilising corpora using traditional and innovative crowdsourcing methods and in co-designing corpus applications (including the Digital Replay System [DRS]).

Almut Koester is a full professor of English business communication at Vienna University of Economics and Business (WU) and, before that, was a senior lecturer in English language at the University of Birmingham in England. She is the author of The Language of Work (2004), Investigating Workplace Discourse (2006) and Workplace Discourse (2010). Her research focuses on spoken workplace discourse and business
corpora, and her publications have examined genre, modality, relational language, vague language, idioms and conflict talk. She is currently exploring the areas of language awareness and creativity and is interested in applying research findings to teaching business English.

**Phoenix Lam** is an assistant professor in the Department of English at The Hong Kong Polytechnic University. She is also a member of the Department’s Research Centre for Professional Communication in English (RCPCE). Her research areas are in corpus linguistics, discourse analysis and intercultural and professional communication. Her publications appear in *Discourse Studies, English for Specific Purposes, Journal of Pragmatics* and *Journal of Sociolinguistics*. Her latest research focuses on the discursive construction of online place branding through a corpus-assisted discourse analytic approach.


**Michaëla Mahlberg** is a professor of corpus linguistics at the University of Birmingham, UK, where she is also the Director of the Centre for Corpus Research. 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, with M. Hoey, M. Stubbs and W. Teubert). She is the editor of the *International Journal of Corpus Linguistics* (John Benjamins) and together with Gavin Brookes she edits the book series *Corpus and Discourse* (Bloomsbury). As Principal Investigator of the AHRC-funded CLiC Dickens project, Michaëla led the development of the CLiC web app.

**Anna Marchi** is a senior assistant professor of English language and linguistics at Bologna University’s Department of Interpreting and Translation. Her research interests include corpus-assisted research methods, news discourse and journalism studies. Recently she has published *Self-reflexive Journalism: A Corpus Study of Journalistic Culture and Community in the Guardian* (Routledge 2019) and co-edited *Corpus Approaches to Discourse: A Critical Review* (Routledge 2018). She earned a PhD at Lancaster University and has worked in the area of corpus linguistics and corpus-assisted discourse studies since 2005 (collaborating with the Universities of Siena, Cardiff, Swansea and Lancaster).

**Geraldine Mark** is a research associate at Cardiff University and PhD student at Mary Immaculate College, Limerick. Her main research interests are in L2 language development, L1 and L2 spoken language, and the use of corpora in informing language teaching and learning. She is co-author of *English Grammar Today* (2011,
Cambridge University Press) and co-Principal Investigator (with Anne O’Keeffe) of the English Grammar Profile, an online resource profiling L2 grammar development.

Gerlinde Mautner is a professor of English business communication at Wirtschaftsuniversität Wien (Vienna University of Economics and Business), visiting professor at Durham Business School, UK, and Vice-President for humanities and social sciences of Austria’s largest research funding organisation (FWF). She pursues research interests located at the interface of language, society, business and the law. Since the mid-1990s, her work has also had a strong methodological focus, concerned in particular with the relationship between corpus linguistics and discourse studies.

Jeanne McCarten taught English in Sweden, France, Malaysia and the UK before starting a publishing career with Cambridge University Press. As a publisher, she had many years’ experience of commissioning and developing ELT materials and was involved in the development of the spoken English sections of the Cambridge International Corpus, including CANCODE. Currently a freelance ELT materials writer, corpus researcher and occasional teacher, her ongoing professional interests lie in successfully applying corpus insights to learning materials. She is co-author of the corpus-informed print, online and blended courses Touchstone and Viewpoint, and Grammar for Business, published by Cambridge University Press.

Michael J. McCarthy is emeritus professor of applied linguistics, University of Nottingham. He is (co)author/(co)editor of 57 books, including Touchstone, Viewpoint, The Cambridge Grammar of English, English Grammar Today, From Corpus to Classroom, Innovations and Challenge in Grammar and titles in the English Vocabulary in Use series. He is author/co-author of 120 academic papers. He was co-founder of the CANCODE and CANBEC spoken English corpora projects. His recent research has focused on spoken grammar. He has taught in the UK, Europe and Asia and has been involved in language teaching and applied linguistics for 56 years.

Tony McEnery is distinguished professor in the Department of English Language and Linguistics at Lancaster University and Changjiang Chair at Xi’an Jiaotong University. He is General Editor of the journal Corpora (Edinburgh University Press) and was founding director of the ESRC-funded Centre for Corpus Approaches to Social Science (CASS) at Lancaster. He has published widely on corpus linguistics and is the author of Corpus Linguistics: Method, Theory and Practice (with Andrew Hardie, Cambridge University Press, 2011) and The Language of Violent Jihad (with Paul Baker and Rachelle Vessey, Cambridge University Press, 2021).

**David Oakey** is a lecturer in TESOL and applied linguistics in the Department of English at the University of Liverpool, UK. His research interest is in the description, acquisition, teaching and learning of lexis, formulaic language and phraseology, particularly in interdisciplinary discourse situations where users encounter unfamiliar meanings of familiar words. He uses corpus linguistics and usage-based approaches to look at the behaviours of forms, meanings and patterns and has published on various aspects of this work. He has taught undergraduate and graduate classes in English lexis and corpus linguistics at universities in China, Turkey, the UK, and the United States.

**Kieran O’Halloran** is a reader in applied linguistics at King’s College, University of London. He is an educationalist who researches and teaches creative thinking and critical thinking using digital technologies. He draws on posthumanist theory for these ends, intersecting his interest in critical thinking with critical discourse studies and an interest in creative thinking with literary stylistics. Publications include *Posthumanism and Deconstructing Arguments: Corpora and Digitally-Driven Critical Analysis* (Routledge 2017) and *Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama* (with David Hoover and Jonathan Culpeper, Routledge 2014).


**Maria Pavesi** is a professor of English language and linguistics at the University of Pavia, where she was Director of the University Language Centre and coordinator of the Foreign Language Teachers’ Education programme. Her research has addressed several topics in English applied linguistics, focussing on informal second language acquisition, the sociolinguistics of dubbing and film dialogue and corpus-based audiovisual translation studies. Her most recent publications include ‘Corpus-Based Audiovisual Translation Studies: Ample Room for Development’ (Routledge 2019) and “‘I Shouldn’t Have Let This Happen”: Demonstratives in Film Dialogue and Film Representation” (Bloomsbury 2020). In 2019, she also co-edited the *Multilingua* special issue on *Audiovisual Translation as Intercultural Mediation*. Since 2005, she has coordinated the development of the *Pavia Corpus of Film Dialogue*, a parallel and comparable corpus of English and Italian film transcriptions.

**Pascual Pérez-Paredes** is a professor of applied linguistics and linguistics, University of Murcia. His main research interests are language variation, the use of corpora in language education and corpus-assisted discourse analysis. He has published research in journals such as *CALL, Discourse & Society, English for Specific Purposes, Journal of Pragmatics, Language, Learning & Technology, System* and the *International Journal of Corpus Linguistics*. He is assistant editor of Cambridge University Press (CUP) *ReCALL* and author of *Corpus Linguistics for Education. A Guide for Research*, in the Routledge Corpus Linguistics Guides series.
Geraint Rees is tenure-track lecturer and Serra Hûnter Fellow at Universitat Rovira i Virgili, Tarragona. His research interests include corpus linguistics, lexicography, vocabulary acquisition and learning technologies. His PhD research, undertaken at Universitat Pompeu Fabra, Barcelona, deals with the selection and presentation of vocabulary in EAP course materials and dictionaries. His current focus is on the integration of lexicographic resources with other technologies. He has worked on ColloCaid, a project developing an integrated text editor and lexicographic resource to help academic writers with English collocations. He is reviews editor for the International Journal of Lexicography.


Ute Römer is an associate professor in the Department of Applied Linguistics and ESL at Georgia State University. Her research interests include phraseology, second language acquisition and using corpora in language learning and teaching. She serves on a range of editorial boards of professional journals and is General Editor of the book series Studies in Corpus Linguistics (John Benjamins). Her research has been published in a range of applied, corpus and cognitive linguistics journals, including Language Learning, The Modern Language Journal, Studies in Second Language Acquisition, Annual Review of Applied Linguistics, International Journal of Corpus Linguistics, Corpora and Cognitive Linguistics.


Tony Berber Sardinha is a professor with the Applied Linguistics Graduate Program, Pontifical Catholic University, Sao Paulo, Brazil. His interests include corpus linguistics, register variation, forensic linguistics, metaphor analysis and applied linguistics. His latest publications include Multi-Dimensional Analysis: Research Methods and Current Issues (Bloomsbury), Multi-Dimensional Analysis: 25 Years On (Benjamins), Metaphor in Specialist Discourse (Benjamins) and Working with Portuguese Corpora (Bloomsbury). He is the Corpus Linguistics editor for the new edition of the Encyclopedia of Applied Linguistics (ed. Carol Chapelle). He serves on the editorial board of major journals in the
field such as the *International Journal of Corpus Linguistics, Corpora, Register Studies* and *Applied Corpus Linguistics*.

**Charlotte Taylor** is a senior lecturer in English language and linguistics at the University of Sussex. She is co-editor of the *CADAAD Journal*. Her current interests lie in the intersection of language and conflict in two contexts: investigating mock politeness and the representation of migration. She also has a keen interest in methodological issues in corpus and discourse work. Her book-length publications include *Corpus Approaches to Discourse* (with Anna Marchi), *Exploring Absence and Silence in Discourse* (with Melani Schroeter), *Patterns and Meanings in Discourse* (with Alan Partington & Alison Duguid) and *Mock Politeness in English and Italian*.

**Ana Maria Terrazas-Calero** is currently a lecturer in English studies at the University of Extremadura (Spain), where she teaches English sociolinguistics, variations of the English language, early modern English and medieval English history. She is also a doctoral student at Mary Immaculate College, University of Limerick (Ireland). She has created the Corpus of Contemporary Fictionalized Irish English as part of her thesis, which investigates the portrayal of this variety in contemporary fiction. Her research and publications focus on pragmalinguistic elements of Irish English, examining their use, pragmatic functions and indexical value regarding contemporary Irish identity as represented in fictional dialogue.

**Paul Thompson** is a reader in applied corpus linguistics at the University of Birmingham, UK. He is currently Acting Director of the Centre for Corpus Research. He is also the founding co-editor of the *Applied Corpus Linguistics* journal (Elsevier). He was a member of the research team on the British Academic Written English (BAWE) and British Academic Spoken English (BASE) corpus projects. His most recent publication is *Interdisciplinary Research Discourse: Corpus Investigations into Environment Journals* (2019, Routledge) co-authored with Susan Hunston.

**Elaine Vaughan** lectures in applied linguistics and TESOL at the University of Limerick, Ireland. She has published on teacher discourse in workplace meetings; communities of practice; the pragmatics of Irish English; humour and laughter; the use of corpora and corpus-based approaches for intra-varietal, pragmatic and sociolinguistic research; corpus-based critical discourse analysis; and media representations of Irish English. She is co-editor, with Carolina Amador-Moreno and Kevin McCafferty, of *Pragmatic Markers in Irish English* (2015, John Benjamins); and with Raymond Hickey of the special issue of the journal *World Englishes* on Irish English (2017).

**Alexandra Vella** holds a PhD in linguistics from the University of Edinburgh. She is an associate professor with the Institute of Linguistics and Language Technology, University of Malta. Her area of expertise is phonetics and phonology, with a focus on prosody. She continues to work on describing the intonational phonology of Maltese and its dialects, also looking to extending insights gained to modelling the intonation of Maltese English, the variety of English of speakers of Maltese. She is the lead researcher on projects involving the compilation and annotation of spoken data of the languages and language varieties of Malta.
Brian Walker is a visiting researcher in the Department of Linguistics and Modern Languages at the University of Huddersfield. His research focusses on the stylistic analysis of texts, including political speeches and print news report, using corpus methods and tools. His recent publications include *Keywords in the Press* (Bloomsbury, 2018; with Lesley Jeffries) and *Corpus Stylistics: Theory and Practice* (Edinburgh University Press, 2019; with Dan McIntyre).

Martin Warren was a professor of English language studies in the English Department of the Hong Kong Polytechnic University and a member of its Research Centre for Professional Communication in English until his retirement in 2018. He taught and researched in the areas of corpus linguistics, discourse analysis, intercultural communication, lexical semantics, phraseological variation, pragmatics and professional communication.

Martin Weisser received his PhD from Lancaster University in 2001 and his professorial qualification (Habilitation) from Bayreuth University in 2011. He is currently an adjunct professor at the University of Bayreuth, specialising in general corpus linguistics, corpus pragmatics and the creation of corpus analysis tools, such as the Dialogue Annotation and Research Tool (DART) and the Simple Corpus Tool. He is the author of *How to Do Corpus Pragmatics on Pragmatically Annotated Data: Speech Acts and Beyond* (Benjamins 2018) and *Practical Corpus Linguistics* (Wiley-Blackwell 2016).

Viola Wiegand is a lecturer at the University of Birmingham. She has contributed to the development of the CLiC web app and research on body language and character speech in nineteenth-century fiction. Her research interests span corpus linguistics and stylistics, the use of corpora in language learning and teaching, discourse analysis, and digital humanities. Viola’s PhD thesis focused on contemporary surveillance discourse in a range of domains. 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).
Acknowledgements

Bringing together this volume for a second time was both an exciting and a challenging endeavour. It would not have been possible without the generosity of the contributors. The compilation of the volume coincided with the COVID-19 global pandemic. This adversely affected society in so many ways. Within the academic community, in addition to personal challenges, various lockdowns brought many practical and logistical difficulties. As we slowly emerge from the gloom of this time, this Handbook, in its small way, is a testament to our collective resilience. To all contributors, including those who had to pull out due to the challenges of this difficult period, we thank you sincerely.

To colleagues and friends in the academic community, we also owe our gratitude. These include, *inter alia*, those who generously acted as reviewers, advisers and conveyers of support and encouragement: Svenja Adolphs, Carolina Amador-Moreno, Laurence Anthony, Tony Berber Sardinha, Gavin Brookes, Graham Burton, Niall Curry, Averil Coxhead, Fiona Farr, Brian Clancy, John Kirk, Dawn Knight, Mike Handford, Michaela Mahlberg, Geraldine Mark, Jeanne McCarten, Tony McEnery, Kieran O'Halloran, Joan O'Sullivan, Pascual Pérez-Paredes, Ute Römer, Christoph Rühlemann, Ana Mª Terrazas-Calero, Paul Thompson, Elaine Vaughan, Martin Warren, Viola Wiegand and Martin Weisser. Sadly, during the writing of this volume, Chris Greaves, one of our contributors, passed away. We also thank our editorial assistant Eleni Steck for her calm help and advice. To Louisa Semlyen, Senior Publisher at Taylor & Francis Group, we thank you for inviting us to edit this volume a second time. We extend a special expression of gratitude to the following doctoral students who provided crucial editorial assistance in this project. Their insights and perspectives were invaluable: Shane Barry, Giana Hennigan, Gerard O’Hanlon, Rose O’Loughlin and Deborah Tobin. Finally, to our partners and family, Ger, Jack and Jeanne, as ever, we thank you sincerely.

Chapter 14 makes use of the *English Vocabulary Profile* and Chapter 22 uses examples from the *English Grammar Profile*. These resources are based on extensive research using the *Cambridge Learner Corpus* and are part of the *English Profile* programme, which aims to provide evidence about language use that helps to produce better language teaching materials. See [http://www.englishprofile.org](http://www.englishprofile.org) for more information.

In Chapter 38, we acknowledge the use of data from films in the *Pavia Corpus of Film Dialogue* (PCFD) in adherence with copyright clearance for use in research publications.
Chapter 47 acknowledges the use of a screenshot from the ‘IBM immersive insights’ YouTube video (Figure 47.1). In the use of Figures 47.4 and 47.5, we acknowledge the permission to use a text from the US Food and Drug Administration website text on animal testing. Additionally, Chapter 47 wishes to acknowledge the permission of Oxford University Press to reuse material from O’Halloran (2020).
1 Corpus linguistics: a decade on

This new edition of the Routledge Handbook of Corpus Linguistics gives us an opportunity to reflect on where the discipline was ten years ago, where it is now and where it might be heading. These three perspectives involve reflection not only on technological advances but also on methodological progress and on the academic and social impact of corpus linguistics (CL).

There can be no doubt that CL has benefited from technological progress in the last decade. Computers work faster; software has, in the main, become more intuitively usable by scholars and others who are non-specialists in computer science; and the mathematical processes which yield results that linguists can interpret have become more sophisticated, and indeed CL has led many of these changes (see Chapters 9 and 13, this volume). The enhanced access to corpora via online interfaces has generally enabled a far broader population to explore data from a greater range of languages than was the case just ten years ago. For example, at the time of writing, the Sketch Engine corpus interface is a repository of over 500 corpora across 95 languages (Kilgarriff et al. 2014). Within this, a user can access the TenTen Corpus Family (so called because their target is to accrue 10 billion or $10^{10}$ words for each language), a suite of comparable corpora comprising collections of web-based texts from over 35 languages. The English-Corpora.org site (curated by Mark Davies and formerly known as BYU Corpora) provides access to multi-million- and multi-billion-word corpora of contemporary and historical English, as well as specialised collections such as The TV Corpus (325 million words), The Movie Corpus (200 million words) and the Corpus of American Soap Operas (100 million words). Clearly in the last decade, the limitations on corpus size have been obviated by the capacity to store vast amounts of data in the cloud but it has also seen the honing of artificial intelligence tools to automatically gather data according to defined curation parameters, thus leading to big data collections that can be rapidly assembled and grown over time.

DOI: 10.4324/9780367076399-1
2 Big data, rapid data, critical perspectives

It is striking, in the last decade, how corpus interfaces can respond rapidly to contemporary themes in society and can curate collections that allow researchers to examine how a phenomenon is both being constructed linguistically and is impacting society. More and more, this is fostering the critical corpus voice and the potential for activism, as evidenced in many works in the last decade, *inter alia*, Grant (2013), Brindle (2018), Dance (2019), Larner (2019), Cunningham and Egbert (2020), Clarke (2019) and Baker *et al.* (2021) (see also Chapters 40–44 and 46–47, this volume).

It has become the norm that major political and societal events come under the corpus linguist’s research gaze. The 2016 Referendum in the UK, for instance, which led to the withdrawal from the European Union (EU) (referred to as Brexit) can be investigated using the *Brexit Corpus* in Sketch Engine. This 100-million-word corpus comprises mostly tweets relating to the referendum, plus news, comments and blogs. A corpus such as this becomes a repository for patterns of and influences on thought: because texts were written before the referendum, researchers can trace opinions and track trends and coinage in language use (see also Islentyeva 2020 on attitudes in the British press to migration amid Brexit).

A more global example of corpus development in rapid response to a major societal event relates to the 2020 COVID-19 global pandemic. At the time of writing (and still amid this pandemic), the *Covid-19 Corpus*, available on Sketch Engine, comprises over 224 million words of texts released as part of the COVID-19 Open Research Dataset (CORD-19). Meanwhile on English-Corpora.org, the *Coronavirus Corpus* makes available over 1 billion words (Davies 2021). The latter corpus was designed to be a record of the social, cultural and economic impact of COVID-19 (Davies 2021). The data give us an insight into shifts in thought about the pandemic. For example, Davies (2021) notes evidence of relative naiveté in March 2020 in the early stages of the virus’s spread where it was viewed as a short-term problem, with the word *re-opening* (of schools and businesses) being noted. One month later, from April 2020 onwards, the corpus shows a shift to a realisation of the longer-term nature of the virus, reflected in the prominence of words and collocates such as *ban + gatherings; travel; sale; entry; mask + wear; wearing; face; required; wore; wears.*

Hyland and Jiang (2021) look at the most highly cited SCI articles on COVID-19 published in the first seven months of 2020 to explore scientists’ use of hyperbolic and promotional language to boost aspects of their research in a quest to ascertain whether this enthusiasm influenced the rhetorical presentation of research and encouraged scientists to “sell” their studies. Their results illustrate a significant increase in hype to stress certainty, contribution, novelty and potential, especially regarding research methods, outcomes and primacy. Hyland and Jiang’s (2021) work sheds light on scientific persuasion at a time of intense social anxiety, and it is a marker of the important role that the linguist fulfils when armed with convincing empirical evidence.

These kinds of large rapid-response corpora, such as the *Brexit Corpus*, the *Covid-19 Corpus* and the *Coronavirus Corpus*, offer an interesting insight into the processes of modern-day corpus-building and neatly illustrate the development since the publication of the first edition of this Handbook one decade ago. For instance, the *Coronavirus Corpus* grows on a daily basis. As Davies (2021) details, new articles come from links harvested from hourly searches of Bing News (as well as over 1,000 websites) to locate articles that have appeared in the previous 24 hours. These are then downloaded, cleaned
of boilerplate material, tagged and lemmatised before being added to the existing corpus. To find articles for the corpus, automatic searches using the search items coronavirus, COVID or COVID-19 are used, plus a list of other words and strings used to search titles, such as at-risk; self-isolat*, lock-down, stockpile*, testing; vaccine; ventilator (see Davies 2021 for a full list). This kind of corpus curation automaticity means we can assume that, as major themes emerge in contemporary society, we are guaranteed to have a corpus with “fresh” data to analyse, and we are much indebted to such endeavours.

Corpora can be a repository of human thought for future analysis. However, there is a need to stop and think about this and to query whether this rapid curation is a “re-fraction” of our shared reality. Now more than ever, such reflection on the creation of big data corpora is crucial because the principles of curation determine how we as corpus creators and analysts represent social reality. O’Halloran in Chapter 47, this volume, brings a very timely and important philosophical perspective to this. O’Halloran’s chapter offers a coda to the Handbook, and it also marks the major change in terms of the proliferation and treatment of data in the last decade (see also O’Halloran 2017).

Another facet of the revolution in mass corpus data gathering has been the use of crowdsourcing. A decade ago, though smart phones were available, one could not assume their widespread use. In recent years, personal mobile devices are at the core of large spoken corpus data-gathering exercises, for example, The National Corpus of Contemporary Welsh (Knight et al. 2021) and The Spoken British National Corpus 2014 (Love et al. 2017). Devices allow for high-quality audio and/or video recordings to be made and easily uploaded, along with consent and user-information documentation. This type of crowdsourcing revolutionises spoken data gathering, but it has implications for the way we design corpora and their sample frames (this is discussed in relation to the British National Corpus 1994 and 2014 in Love et al. 2017 and Love 2020).

3 Many text types, many modes

What is considered corpus data is also changing. Corpus content used to be heavily focused on breadth of representation of scanned written texts, e.g. the 100-million-word written text component of the British National Corpus (BNC1994). As discussed, the use of personal devices and automated tools to harvest data means that a corpus can now be whipped up in short order. Anyone interested in building a corpus on a particular topic can do so much more quickly than a decade ago and can access a far wider range of text types.

Early corpora were slow and considered enterprises. Design matrices were refined as sampling frames (see Crowdy 1993; McCarthy 1998). Now, with so many data types available in electronic form, the capacity to automatically find, harvest and store these has never been easier. However, there is still a need to keep in mind the original principles of careful corpus curation and design. Documenting design decisions and rationales remains crucial, and work such as Love (2020), which gives insights into the design decisions and rationales of the Spoken BNC 2014, is essential if we are to offset the risk of taking our eye off long-held principles. Amid so much change in terms of what constitutes “data”, have we as a community brought up to date what it means to create a representative corpus in terms of text types? Are binary terms like “spoken” and “written” corpora any longer fit for purpose? The proliferation of new media since the publication of the first volume of this Handbook causes us to reflect on these questions.
At the time of its publication in 2010, Facebook, Twitter and online content streaming platforms such as YouTube were all in their start-up phase. Netflix had just begun its streaming enterprise, and virtual communication via computer-mediated communication tools such as Skype was not widely adopted. WhatsApp, Facetime, Instagram, Zoom and TikTok (originally Musical.ly) did not launch until 2009, 2010, 2010, 2013 and 2014, respectively, and did not become widely used until much later as mobile personal devices, especially smart phones, became more popular. From a corpus linguistics perspective, all of these applications and platforms have already had an impact on the creation, definition and processing of data. These data are easy pickings for corpus builders because they are very “captureable”, but, we ask, do we have the capacity to treat these data for what they are: multi-modal? Chapters 40 and 46 on corpora of news and social media are new chapters in this edition of the Handbook, and they bring a timely perspective to this issue, as does Chapter 7.

In the introductory chapter to the first edition of this Handbook in 2010, we expressed our excitement about being at the point where technology allowed for ‘the creation of multi-modal corpora, in which various communicative modes (e.g. speech, body-language, writing) could all be part of the corpus, all linked by simple technologies such as time-stamping and all accessible at one go’ (McCarthy and O’Keeffe 2010: 6). We proclaimed that those interested in the study of spoken language using corpus linguistics would no longer ‘have to rely only on the transcript of a speech event’ because we are now at a point where video and audio stream can be tied to the transcript, thus ‘offering invaluable contextual and para-linguistic and extra-linguistic support to the analysis’ (2010: 6) (see also Knight and Adolphs 2020). Alas, while we were factually accurate, we were very premature in our excitement. A decade on, we are still working with impoverished representations of spoken language. What is more acute now, however, as mentioned, is that we have never before had so much multi-modal content, from online streaming to social media. Closed-caption facilities available on many content sharing and communication platforms are of a relatively high standard and thus lessen the chore of transcription, or at least give the researcher a major head start. However, while we are surrounded by multi-modal data, we are still working with a text-based paradigm for their investigation. Our default setting is to reduce all data to a text-based representation of discourse, regardless of the richness of its provenance. This is a pressing issue for the study of social media content where so often intertextuality is central to the message. Our systems and protocols are still not ready for these new data, and we need to find ways to capture and process multi-modal features (e.g. how can we best incorporate emojis; gifs; or the combination of audio, video, emoji, text and filters in a TikTok?).

Treating rich multi-modal data as a reduced one-dimensional written transcript devoid of accompanying sound, image, gesture, gaze, head nods, etc., means, as Rühlemann (2019) notes, we are observing only the transcribed speech of a given communicative event rather than the unified whole. The technical, financial and temporal challenges of transcribing and capturing the entire multi-modal “bundle” (Crystal 1969) means that researchers continue to make do with impoverished orthographic transcription of what has been said at the expense of how it has been communicated. We would argue that, as a research community, the lack of widespread advances in multi-modal transcription has caused a stagnation and has meant that we are largely ignoring both the richness of the data and the affordances of the digital technology that are currently available to us. In this edition of the Handbook, we make no predictions about what will happen in the next decade in this respect, but we do note our desire for change.
4 The impact of learner corpus research

Learner corpora were traditionally designed, built and analysed by researchers as sources of information on areas such as contrastive error analysis (L1 versus L2 interference). Many studies led to better understandings of learner language as a system (see Granger et al. 2015). Interestingly in the last decade, the contrastive focus is shifting. More and more, we are learning about what learners of a language can do. We are seeing a profiling turn where, very much influenced by calibrated proficiency scales like the Common European Framework of Reference for Languages (CEFR, Council of Europe 2001), there is a need to empirically test these competency profiles (e.g. Thewissen 2013; O’Keeffe and Mark 2017; McEnery et al. 2019; see also Chapters 22 and 23 this volume).

Using corpora to test intuitively derived competence descriptors and scales for what constitutes a given level of competence in a language is an area ripe for further development. Profiling learner competence has also been moving apace in machine learning. More and more learner data are fed to machines as training data. The learner data thus become fodder for the machine, which is being trained to process learner performances in terms of group or individual profiling for purposes of adaptive feedback or automated assessment. The Institute for Automated Language Teaching and Assessment (ALTA) at the University of Cambridge, for example, develops methods for operations such as automatic speech recognition, learner profiling, automated essay grading and automated feedback systems based around learner corpora, applying novel techniques of language modelling to the data (for example, Felice et al. 2014). Such approaches include three-dimensional data cubing, where learners occupy one dimension, their disciplines the second and change over time the third, on the basis of which individuals and groups can be “modelled” for machines to take over the work of analysis and feedback. Similarly at ALTA, the development of a chatroom corpus of teacher–learner interactions is targeted at creating a chatbot to give adaptive feedback independently, but closely aligned with, typical teacher–student feedback.

Notable also in the last decade, and reflected in Chapters 22 and 23, among others, in this volume, is the move towards the use of CL in pursuit of second language acquisition (SLA) research questions. This is long overdue, as noted by Johansson (2009); De Knop and Meunier (2015); Flowerdew (2015); McEnery et al. (2019) and O’Keeffe (2021), among others. Corpus-based work on usage-based perspectives on SLA is growing rapidly, building on seminal work such as Ellis et al. (2015). However, as Römer and Garner note (Chapter 23, this volume), echoing Mackey (2014), partnerships between experts on SLA and CL need to be fostered if the benefits of longitudinal learner corpora are to be fully reaped. O’Keeffe (2021) also notes the need for reaching out to SLA for an enhanced understanding of and rationale for data-driven learning (DDL) (see also Chapters 21 to 33, this volume).

5 Our history and our future

In our introduction to the first edition of this Handbook, we said that CL was perhaps most readily associated in the minds of linguists with searching through screen after screen of concordance lines and wordlists in an attempt to make sense of phenomena in texts. We noted that this method of exegesis can be traced back to the thirteenth century, when biblical scholars and their teams of minions pored over the Christian Bible and manually indexed its words, line by line, page by page. Concordancing arose out of a
practical need to specify for other biblical scholars, in alphabetical arrangement, the
words contained in the Bible, along with citations of where and in what passages they
occurred. In 2021, concordancing programmes can replicate the work of 500 monks in
nanoseconds. Though concordances go back centuries as a laborious practice, they re-
present a significant link to past scholarship, and their spirit remains at the heart of the
discipline for any linguist hoping to go beyond the number crunching of frequency lists,
keyword lists, cluster lists and so on. We are still, in the final analysis, practitioners of
rhetoric. The persuasiveness of our arguments about language depends on the plausible
and robust interpretation of the principled empirical evidence which the data throw up.
Grammatical description is a case in point, and Halliday’s position rings as true today as
a quarter of a century ago:

… the corpus does not write the grammar for us. Descriptive categories do not
emerge out of the data. Description is a theoretical activity; and … a theory is a
designed semiotic system, designed so that we can explain the processes being ob-
erved.

(Halliday 1996: 24)

Our debt to the past goes beyond an acknowledgement of the laborious process of
manual concordancing of sacred texts and can be extended to the work of lexicographers
and that of pre-Chomskyan structural linguists. In both cases, collecting reliable data
was essential to their work. Dr Samuel Johnson’s first comprehensive dictionary of
English, published in 1755, was the result of many years of working with a corpus of
endless slips of paper, logging samples of usage from the period 1560 to 1660. And
perhaps the most famous example of the ‘corpus on slips of paper’ were the more than 3
million slips attesting word usage that the Oxford English Dictionary (OED) project had
amassed by the 1880s, stored in what nowadays might serve as a garden shed. These
millions of bits of paper were, quite literally, pigeon-holed in an attempt to organise
them into a meaningful body of text from which the world-famous dictionary could be
compiled.

The first computer-generated concordances had appeared in the late 1950s, using
punched-card technology for storage (see Parrish 1962 for an early discussion of the
issues). At that time, the processing of some 60,000 words took more than 24 hours.
However, considerable improvements came about in the 1970s. Meanwhile, from as
early as 1970, library and information scientists had developed a keen interest in Key
Word In Context (KWIC) concordances as a way of replacing catalogue indexing cards
and of automating subject analysis (Hines et al. 1970), and many well-known biblio-
ographies and citation source works benefitted from advances in computer technology.
Before it found its way into the linguistic terminology, the term corpus had long been in
use to refer to a collection or binding together of written works of a similar nature. The
OED attests its use in this meaning in the eighteenth century, such that scholars might
refer to a ‘corpus of the Latin poets’ or a ‘corpus of the law’. The OED’s first citation of
the word corpus in the linguistic literature is dated at 1956, in an article by W. S. Allen in
the Transactions of the Philological Society, where it is used in the more familiar meaning
of ‘the body of written or spoken material upon which a linguistic analysis is based’
(OED online, 2021). McNerney et al. (2006) note that the more specific term corpus lin-
guistics did not come into common usage until the early 1980s; Aarts and Meijs’s work
(1984) is seen as the defining publication with regard to coinage of the term.
As editors, we have had the privilege of bringing this volume together for a second time. Like a time capsule, it may be opened in ten years’ time, and a decade may again show stunning changes and it may still point to a lack of progress in other respects. This has clearly been a decade of immense growth for corpora. We have such a richness of ever-growing data and ever-advancing tools. In summary, there is little to limit the timeliness and the bigness of corpora now, but with this comes both a risk and a responsibility to safeguard the tenets of principled sampling, corpus design and representativeness. We hope this Handbook can play a role in this process.

We have structured the 46 main chapters of the volume under four themes:

- Building and designing a corpus: the basics;
- Using a corpus to investigate language;
- Corpora, language pedagogy and language acquisition;
- Corpora and applied research.

Any one of these themes could have filled an entire handbook in its own right, but we hope that the selection of chapters that we bring together continues in the spirit of the first Routledge Handbook of Corpus Linguistics to enthuse anyone wishing to get started with building and analysing a corpus. We hope it will inspire readers to continue to build more corpora (big and small), to use corpora more to enhance language learning and to mine more corpora in the pursuit of robust critical perspectives on the use of language and discourse in the construction of our shared reality.

**Note**

1 From *Sailing to Byzantium*, a poem by William Butler Yeats, first published in 1928.

**References**


'Of what is past, or passing, or to come': corpus linguistics, changes and challenges


Building a corpus: what are key considerations?


Biber, D. , Conrad, S. and Reppen, R. (1998) Corpus Linguistics: Exploring Language Structure and Use, Cambridge: Cambridge University Press. (This book provides an overview of corpus linguistics and its many applications, including discovering patterns of language use to researching language change over time. The chapters build from the lexical to the discourse level, each with detailed examples of studies related to the topic being covered in the chapter. The book ends with a series of methodology boxes that provide readers with answers to many of the methodological processes related to using corpora for research.)


O'Keeffe, A. , McCarthy, M. J. and Carter, R. A. (2007) From Corpus to Classroom: Language Use and Language Teaching, Cambridge: Cambridge University Press. (The authors have done extensive research on language and patterns of use. This information is the foundation for the practical applications of corpus research that is presented to English-language teachers. In addition to English-language teachers, language researchers will see this book as a wonderful resource on many aspects of language, especially spoken language.)

Reppen, R. and Simpson-Vlach, R. (2020) Corpus Linguistics, in N. Schmitt and M. Rodgers (eds) An Introduction to Applied Linguistics, London: Arnold, pp. 91–108. (This chapter presents an overview of corpus linguistics and highlights how the methodology of corpus linguistics can be used to explore many areas of interest in the area of applied linguistics.)

Laurence Anthony's homepage https://www.laurenceanthony.net/ (This has links to a variety of resources for both analysing and building corpora. In addition to the well-known AntConc concordancing software, there is AntCoreGen (2019) that allows users to build discipline-specific corpora. This site also includes tools for converting files, analysing vocabulary, annotating texts for part of speech (PoS) and analysing n-grams).

CROW (Corpus repository of writing) writecrow.org (This site has extensive resources that include a web interface with a corpus of university student writing in English and the assignments used to generate the writing. It also has links and free resources to help researchers create and analyse corpora.)

Building a spoken corpus: what are the basics?


ioeconomicclassificationsssecurebasedonsoc2010 [Accessed 31 March 2021].

Building a written corpus: what are the basics?

Building small specialised corpora

Flowerdew, L. (2004) The Argument for Using English Specialized Corpora to Understand Academic and Professional Settings, in U. Connor and T. Upton (eds) Discourse in the Professions, Amsterdam: John Benjamins, pp. 11–33. (This chapter is useful for anyone wanting to build a specialised corpus. As well as presenting a rationale for using specialised corpora, it provides useful guidelines for defining a specialised corpus and for corpus design.)

Handford, M. (2010) The Language of Business Meetings, Cambridge: Cambridge University Press. (This book provides a complete description of all the steps involved in building and exploiting a corpus of one professional genre (the business meeting) from data collection and corpus compilation to corpus analysis and interpretation.)

corpus linguistics can be combined with other methods, in this case ethnography and conversation analysis.)

O’Keeffe, A., McCarthy, M. J. and Carter, R. A. (2007) From Corpus to Classroom: Language Use and Language Teaching, Cambridge: Cambridge University Press. (This book provides an accessible introduction to the most important topics in corpus research. The role of qualitative as well as quantitative analysis is a theme throughout the book, and many chapters address the topic of what can be learned from small, specialised corpora, in particular Chapters 8 and 10.)


BASE (British Academic Spoken English) and BASE Plus Collections, https://warwick.ac.uk/fac/soc/al/research/collections/base/ [Accessed 9 August 2020].


Building a corpus to represent a variety of a language

Biber, D. (1993) Representativeness in Corpus Design, Literary and Linguistic Computing 8(4): 243–257. (In this article, Biber outlines how to construct a statistically representative corpus. In common with many seminal texts, this work has proven to be both inspirational and contentious.)


Kučera, K. (2002) The Czech National Corpus: Principles, Design, and Results, Literary and Linguistic Computing 17(2): 245–257. (Both of these articles explore, and discuss practical solutions to, the problems encountered during the design and construction of large, non-English language corpora representative of language varieties.)

Meyer, C. (2002) English Corpus Linguistics: An Introduction, Cambridge: Cambridge University Press. (This book provides an accessible introduction to corpus linguistics in addition to a step-by-step guide to corpus design, construction and analysis. Meyer draws heavily on corpora such as the BNC and ICE in order to illustrate each stage.)


Joyce, P. W. (1910) English as We Speak it in Ireland, Dublin: M.H. Gill.


Building a specialised audiovisual corpus


What corpora are available?


**What can corpus software do?**


What are the basics of analysing a corpus?
Adolphs, S. and Carter, R. A. (2013) Spoken Corpus Linguistics: From Monomodal to Multimodal, London: Routledge. (This provides a useful description and analysis, showing some of the possibilities available when designing and working with spoken corpora.)
Collins, L.C. (2019) Corpus Linguistics for Online Communication: A Guide for Research, London: Routledge. (This is a useful and highly practical introduction to using corpora to investigate forms of online communication such as the use of social media.)
Hoey, M. (2005) Lexical Priming: A New Theory of Words and Language, London: Routledge. (This is an influential book, which shows how corpora can be used to investigate and further our understanding of language in use.)
Timmis, I. (2015) Corpus Linguistics for ELT: Research and Practice, London: Routledge. (Aimed at those teaching English as a second or foreign language, this is a useful and practical introduction to corpus linguistics which will benefit anybody interested in working with corpora.)
How can a corpus be used to explore patterns?

Hunston, S. and Perek, F. (eds) (2019) Constructions in Applied Linguistics, special issue of the International Journal of Corpus Linguistics 24(3). (This journal issue explores the applicability of patterns to discourse analysis and language teaching, the development of resources such as constructicon and the connections between patterns and construction grammar. It comprises papers by U. Römer, S. Gries, N. Groom, S. Hunston and F. Perek and A. Patten.)


McSorley, E. and Patten, A. (2019) Addressing the Vocabulary Gap Using the Pattern Grammar Approach, Impact 6 xx. (This paper discusses how attention to patterns can improve the teaching of academic vocabulary.)


What can corpus software reveal about language development?


Lu, X. (2014) Computational Methods for Corpus Annotation and Analysis, Singapore: Springer. (This book provides a systematic and accessible introduction to diverse types of computational tools that can be used for automatic or computer-assisted annotation and analysis of text corpora at various linguistic levels.)

MacWhinney, B. (2000) The CHILDES Project: Tools for Analyzing Talk, 3rd edn, Mahwah: Lawrence Erlbaum Associates. (This book provides hands-on instruction on how to transcribe naturalistic child language development data following the CHILDES format and automatically analyse such data using CLAN. Readers are introduced to a set of computational tools designed to improve the readability of transcripts, to automate the data analysis process and to facilitate the sharing of transcribed data.)


VanPatten, B. and Williams, J. (eds) (2014) Theories in Second Language Acquisition: An Introduction, 2nd edn, New York: Routledge. (This edited volume presents a comprehensive introduction to early and contemporary theories in second language acquisition. It provides an excellent overview of each of these compelling theories.)


How to use statistics in quantitative corpus analysis

Gries, St. Th. (2013) Statistics for Linguistics with R, 2nd edn, Berlin and Boston: De Gruyter Mouton. (A still very useful overview of statistical methods, focusing mostly on corpus data and different kinds of regression modeling, but discussing also hierarchical cluster analysis.)

Levshina, N. (2015) How to Do Linguistics with R: Data Exploration and Statistical Analysis, Amsterdam: John Benjamins. (A textbook on R in linguistics in general, with many applications pertinent to the sections in this chapter.)

Paquot, M. and Gries, St. Th. (eds) (2021) Practical Handbook of Corpus Linguistics, Berlin and New York: Springer. (A new handbook of corpus linguistics with overview chapters on many central corpus-linguistic notions, as well as many hands-on chapters on statistical techniques applied to corpus data using R.)


What can a corpus tell us about lexis?

Deignan, A. H. (2017) From Linguistic to Conceptual Metaphors, in E. Semino and Z. Demjen (eds) The Routledge Handbook of Metaphor and Language, London: Routledge, pp. 102–117. (This chapter is a very readable 'way in' to using corpora to study metaphor and makes a convincing case for the necessity of using corpus data to inform research into this important area of lexis at the interface of E-language and I-language.)


Hanks, P. (2013) Lexical Analysis, Cambridge, MA: MIT Press. (This book propounds a lexically driven theory of language reflecting the tendency of users to choose certain ways of expressing themselves. A fascinating overview of words and meanings and how their use changes over time.)

Hasselgård, H., Ebeling, J, and Oksefjell Ebeling, S. (eds) (2013) Corpus Perspectives on Patterns of Lexis, Amsterdam: John Benjamins. (This collection of papers illustrates pertinent questions of lexis that can be investigated by a corpus linguistic approach.)

Murphy, M. L. (2010) Lexical Meaning, Cambridge: Cambridge University Press. (This book is a readable survey of traditional concerns in the study of lexis and is useful for benchmarking corpus linguistic studies.)


What can a corpus tell us about multi-word units?


What can a corpus tell us about grammar?

Biber, D. , Johansson, S. , Leech, G. , Conrad, S. and Finegan, E. (2021) Grammar of Spoken and Written English, Amsterdam: John Benjamins. (This reference grammar covers frequency information, lexico-grammar patterns and comparisons of use in conversation, fiction writing, newspaper writing and academic prose for the major structures in English, in addition to chapters on fixed phrases, stance and conversation. It includes everything from the earlier Longman Grammar of Spoken and Written English.)

Carter, R. A. and McCarthy, M. J. (2006) Cambridge Grammar of English, Cambridge: Cambridge University Press. (This reference grammar emphasises spoken vs. written language for major grammar features of English and includes many lexico-grammatical analyses. It also covers some functional categories, such as typical grammatical realisations of speech acts and many typical ESL difficulties.)


What can a corpus tell us about registers and genres?


Römer, U., Cortes, V. and Friginal, E. (eds) (2020) Advances in Corpus-Based Research on Academic Writing: Effects of Discipline, Register, and Writer Expertise, Amsterdam: John Benjamins. (This edited volume assembles 14 corpus-based studies of a range of written academic registers (EFL student writing, first-year composition, research articles across disciplines, stand-alone-literature reviews, conference abstracts). Studies also demonstrate a range of linguistic foci: academic vocabulary, lexical bundles and p-frames, verb constructions, adjectives as nominal pre-modifiers and multi-dimensional analyses.)


What can a corpus tell us about discourse?


Collins, L. C. (2019) Corpus Linguistics for Online Communication: A Guide for Research, London: Routledge. (Although the book’s focus is on online communication, it is also instructive for those doing corpus-assisted research on discourse in other areas. It also contains a glossary and tasks, with commentaries in an appendix.)

Taylor, C. and Marchi, A. (eds) (2018) Corpus Approaches to Discourse: A Critical Review, London: Routledge. (This edited volume is a thought-provoking account that experienced researchers are likely to find particularly useful. In three parts, it examines hitherto overlooked areas, triangulation and questions of research design.)


**What can a corpus tell us about pragmatics?**


Grice, H. P. (1975) Logic and Conversation, in P. Cole and J. L. Morgan (eds) Syntax and Semantics III, New York: Academic Press, pp. 43–58. (This paper is perhaps one of the most fundamental works in pragmatics.)


What can a corpus tell us about phonetic and phonological variation?

Durand, J., Gut, U. and Kristoffersen, G. (eds) (2019) The Oxford Handbook of Corpus Phonology, Oxford: Oxford University Press. (A good place to go for an in-depth understanding of phonological analysis with a basis in CL. Delais-Roussarie and Post [Chapter 4] discuss issues also raised in this chapter. Also dealt with are applied studies in CL and phonology, methodology and key corpora designed for doing phonology.)


5 in particular discusses variation as a language-internal phenomenon, situating it within the speech community, whilst also dealing with sound change.)


What can a corpus tell us about language teaching?


Davies, M. (2020) corpus.byu.edu, Available at: https://corpus.byu.edu/overview.asp.
What can corpora tell us about language learning?


What can corpus linguistics tell us about second language acquisition?


Hasko, V. (2013) Capturing the Dynamics of Second Language Development via Learner Corpus Research: A Very Long Engagement, The Modern Language Journal 97(S1): 1–10. (This introduction to the journal's special issue discusses the potential of learner corpora for SLA and some of the challenges that need to be overcome if we want to bring the fields of learner corpus research and SLA closer together.)

McEnery, T. , Brezina, V. , Gablasova, D. and Banerjee, J. (2019) Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyse Language Use, Annual Review of Applied Linguistics 39: 74–92. (This position paper discusses the brief history of learner corpus research, its (so far limited) engagement with SLA and opportunities and challenges for future collaboration between the two fields.)

Meunier, F. (2015) Developmental Patterns in Learner Corpora, in S. Granger , G. Gilquin and F. Meunier (eds) The Cambridge Handbook of Learner Corpus Research, Cambridge: Cambridge University Press, pp. 379–400. (This handbook chapter reviews some of the core issues regarding the use of learner corpora to investigate developmental patterns in SLA, including study design and methods for analysing group and individual trajectories. It also describes future avenues for learner corpus design that allow for better analysis of developmental trajectories.)


Römer, U. and Berger, C. M. (2019) Observing the Emergence of Constructional Knowledge: Verb Patterns in German and Spanish learners of English at Different Proficiency Levels, Studies in Second Language Acquisition, Advance online publication. 10.1017/S0272263119000202


What can a corpus tell us about vocabulary teaching materials?

Cortes, V. (2006) Teaching Lexical Bundles in the Disciplines: An Example from a Writing Intensive History Class, Linguistics and Education 17: 391–406. (This study describes a teaching approach to lexical bundles in a writing history class which involved pre- and post-instruction of lexical bundles. The students were exposed to the data in a corpus of history journal articles where four-word lexical bundles were identified.)


Nekrasova-Beker, T., Beker, T. and Sharpe, A. (2019) Identifying and teaching Target Vocabulary in an ESP Course, TESOL Quarterly 10(1): 1–27. (Based on a case study from an introductory engineering course, this study describes the procedure to develop teaching materials for L2 vocabulary instruction using corpus-based techniques.)


What can a corpus tell us about grammar teaching materials?


Gablasova, D., Brezina, V. and McEnery, T. (2019) The Trinity Lancaster Corpus: Applications in Language Teaching and Materials Development, in S. Götz and J. Mukherjee (eds) Learner Corpora and Language Teaching, Amsterdam: John Benjamins, pp. 8–28. (This chapter discusses how the TLC can be applied to language teaching and the development of teaching materials to develop speaking skills, focussing particularly on expression of disagreement, ability to adjust language choice according to linguistic setting, and engaged listenership.)

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. (This paper explains how the EGP was created. This is interesting in that it shows how it is possible to move from raw, learner corpus data to a pedagogical resource for materials designers and teachers.)

Timmis, I. (2015) Corpus Linguistics for ELT: Research and Practice, London: Routledge. (Timmis’ book offers a comprehensive account of how corpora can be applied to ELT. There is an entire chapter on corpus research and grammar, and the discussion at times offers an interesting counterpoint to Biber and Reppen (2002) above, explaining how pedagogical arguments might sometimes be more relevant than frequency-based arguments.)


Corpus-informed course design


Using corpora to write dictionaries

RAE (no date) Banco de datos, RAE, Available at https://www.rae.es/recursos/banco-de-datos [Accessed 16 April 2020].

What can corpora tell us about English for Academic Purposes?

Anthony, L. (2019) Tools and Strategies for Data-Driven Learning (DDL), in K. Hyland and L. Wong (ed.) Specialised English: New directions in ESP and EAP Research and Practice, London: Routledge, pp. 162–180. (This chapter provides a useful and readable discussion of DDL and a range of tools and strategies. While the focus of the chapter is English for specific purposes (ESP), much of this chapter is just as relevant to EAP.)
Biber, D. (2006) University Language: A Corpus-Based Study of Spoken and Written Registers, Amsterdam: John Benjamins. (This book is fundamental reading for EAP researchers and teachers, as it provides an in-depth and thoughtful analysis of what corpora can show us about EAP.)


What is data-driven learning?


Lenko-Szymanska, A. and Boulton, A. (eds) (2015) Multiple Affordances of Language Corpora for Data-Driven Learning. Amsterdam: Benjamins. (This volume includes introductory chapters relating corpora to language pedagogy and language-learning theories, followed by chapters reporting on projects involving corpora in both language learning and translator education.)

Poole, R. (2018) A Guide to Using Corpora for English Language Learners, Edinburgh: Edinburgh University Press. (This book is aimed at both teachers and learners. It is a practical guide which provides suggestions for learners and teachers who wish to incorporate consultation of corpus data in the language-learning process.)


Using data-driven learning in language teaching


Leńko-Szymańska, A. and Boulton, A. (eds) (2015) Multiple Affordances of Language Corpora for Data-Driven Learning, Amsterdam: John Benjamins. (This volume presents advances in DDL, based on corpora for language learning, skills development and translation training and using a variety of research methodologies.)

O'Keeffe, A. (2021) Data-Driven Learning – A Call for a Broader Research Gaze, Language Teaching 54(2): 259–272. (This paper provides an innovative perspective by showing how DDL can be positioned theoretically. It also identifies avenues for further research in DDL.)

Poole, R. (2018) A Guide to Using Corpora for English Language Learners, Edinburgh: Edinburgh University Press. (Designed for both language learners and language teachers, this book provides illustrated examples of how to search corpora, interpret findings and build one's own corpus; it includes many screenshots and hands-on exercises.)


Using corpora for writing instruction

Charles, M. and Frankenberg-Garcia, A. (eds) (2021) Corpora in ESP/EAP Writing Instruction, London: Routledge. (This edited volume deals with the following areas: corpus use for preparing DDL instruction, corpus use by students and corpus use for analysing student writing.)


Karpentko-Seccombe, T. (2020) Academic Writing with Corpora, London: Routledge. (This book provides a practical introduction to DDL, using Lextutor as well as other resources such as SkELL and MiCUSP.)


BAWE Quicklinks (n.d.). Available at: https://bawequicklinks.coventry.domains/ [Accessed 21 September 2020].


Peter Lang, pp. 15–45.

How can corpora be used in teacher education?
Farr, F. and O'Keeffe, A. (2019) Using Corpus Approaches in English Language Teacher Education, in S. Walsh and S. Mann (eds) The Routledge Handbook of English Language Teacher Education, London: Routledge, pp. 268–282. (This outlines a number of ways in which localised and specialised corpora in particular can be included to enhance the ways in which teacher education is conducted.)
McCarthy, M. J. (2008) Accessing and Interpreting Corpus Information in the Teacher Education Context, Language Teaching 41(4): 563–574. (This raises issues in relation to the lack of inclusion of corpus-based approaches and materials in teacher education and outlines the ways in which they could and should be included to give future teachers more critical adaptation skills and influential lobbying power with researchers and publishers of corpus-based teaching materials.)
Vásquez, C. and Reppen, R. (2007) Transforming Practice: Changing Patterns of Participation in Post-Observation Meetings, Language Awareness 16(3): 153–172. (This illustrates how the collection of a corpus of teaching practice feedback meetings was used to identify what was considered to be an unsatisfactory participation balance between students teachers and tutors, the remedial action taken by the tutors and the resulting improvement in the second part of the teaching practice cycle.)


Vaughan, E. (2007) 'I think we should just accept ... our horrible lowly status': Analysing Teacher-Teacher Talk within the Context of Community of Practice, Language Awareness 16(3): 173–188.

How can teachers use a corpus for their own research?

O'Keeffe, A., McCarthy, M. J. and Carter, R. A. (2007) From Corpus to Classroom: Language Use and Language Teaching, Cambridge: Cambridge University Press. (This practical introduction to the discipline of corpus linguistics remains one of the most popular for good reason. It provides an extensive overview of the relationship between corpus-based research and language teaching, using numerous practical examples of the types of linguistic information a corpus can provide.)

Timmis, I. (2015) Corpus Linguistics for ELT, London: Routledge. (This book surveys the relationship between research in corpus linguistics and practice in ELT critically and in depth. It advocates for greater practitioner awareness and engagement vis-a-vis corpus research and asserts this as a critical factor for the reflective professional.)


O’Keeffe, A. (2020) Data-Driven Learning – A Call for a Broader Research Gaze, Language Teaching 1–14. 10.1017/S0261444820000245


How to use corpora for translation

Beeby, A., Rodríguez Inés, P. and Sánchez-Gijón, P. (eds) (2009) Corpus Use and Translating. Corpus Use for Learning to Translate and Learning Corpus Use to Translate, Amsterdam: Benjamins. (This volume is a collection of papers on different aspects of corpus use in the classroom, including reports on corpus use by learners, corpus construction, use of specialised and general corpora and their use for evaluation purposes. It will be of interest to translator trainers and trainees and researchers in applied linguistics, corpus linguistics and translation studies.)
Zanettin, F. (2012) Translation-Driven Corpora. Corpus Resources for Descriptive and Applied Translation Studies, Oxford: Routledge. (This handbook covers corpus design, encoding and analysis, with a special focus on multilingual corpora and translation-oriented research questions, providing extensive exemplification and activities.)
Written English, London and New York: Longman.
University Press.
Training Purposes in N. Kübler (ed.) Corpora, Language, Teaching and Resources: From Theory to Practice,
Bern: Peter Lang, pp 221–248.
assisted Case Study in Translating Evaluative Prosody’, in A. Ferraresi, R. Pederzoli, R. Scansani and S.
Cavalcanti (eds) Mediazioni Special Issue on Research Methods and Themes in Translation, Interpreting
and Intercultural Studies 29, online: http://www.mediazioni.sitlec.unibo.it/index.php/no-29-special-issue-
2020.html.
by a Group of 13 Students at a UK University, Corpora 10(2), online: https://www.euppublishing.com/doi/full/10.3366/cor.2015.0081.
Corpus, in D. Stewart, F. Zanettin and S. Bernardini (eds) Corpora in Translator Education, Manchester: St
and I. Delaere (eds) Empirical Translation Studies New Methodological and Theoretical Traditions, Berlin: De
Gruyter Mouton, pp. 9–46.
Amsterdam: Benjamins.
487–500.
Kilgarriff, A., Baisa, V., Buštá, J., Jakubiček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V.
Contacts, Translation and Second Language Acquisition, in H. Paulasto, L. Meriläänen, H. Rionheimo and
M. Kok (eds) Language Contacts at the Crossroads of Disciplines, Newcastle: Cambridge Scholars
Publishing, pp. 3–32.
Lanstyák, I. and Heltai, P. (2012) Universals in Language Contact and Translation, Across Languages and
and Technical Translation training, JoSTrans. The Journal of Specialised Translation 9, online:


---

**Using corpus linguistics to explore the language of poetry: a stylometric approach to Yeats’ poems**

Hoover, D., Culpeper, J. and O’Halloran, K. (2014) *Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama*, London: Routledge. (This book covers both corpus and computational approaches to stylistics and includes numerous case studies exemplifying some of the techniques discussed in this chapter.)

the corpus stylistic analysis of poetry particularly.)
Murphy, S. (2015) I Will Proclaim Myself What I Am: Corpus Stylistics and the Language of Shakespeare's Soliloquies, Language and Literature 24(4): 338–354. (This article demonstrates the value of keyword analysis in differentiating texts according to genre.)
Anthony, L. (2019) AntConc 3.5.8, Tokyo, Japan: Waseda University, Available at: https://www.laurenceanthony.net/software.
Using corpus linguistics to explore literary speech representation: non-standard language in fiction

Adolphs, S. (2006) Introducing Electronic Text Analysis: A Practical Guide for Language and Literary Studies, London: Routledge. (This is a good introduction to the topic of written corpora, and it provides very useful insights into the analysis of electronic texts in general.)

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. (This paper analyses keywords in *Romeo and Juliet*. It is a good introduction to corpus analysis in the context of drama. Keyword analysis, as the author suggests, can be applied to other kinds of data, such as particular registers, dialects, media, documents or writings.)

McIntyre, D. and Walker, B. (2019) Corpus Stylistics: Theory and Practice, Edinburgh: Edinburgh University Press. (This is the most updated monograph explaining the theory and methodology of corpus stylistics with practical examples. It begins with an introduction to the theory of corpus linguistics and stylistics and how they intersect. It also offers an overview of the current state of the field and explores the application of corpus stylistics methods to pedagogy, historical linguistics and even its potential use in the real world.)


Exploring narrative fiction: corpora and digital humanities projects

Mahlberg, M. (2013) Corpus Stylistics and Dickens’s Fiction, London: Routledge. (This monograph introduces a lexically driven approach to the building blocks of fictional worlds. It combines quantitative and qualitative analyses of clusters, i.e. repeated sequences of words, and studies their textual functions for characterisation. Beyond its particular focus on Dickens, the book outlines fundamental principles of corpus stylistics.)

Mastropierro, L. (2017) Corpus Stylistics in Heart of Darkness and its Italian Translations, London: Bloomsbury. (Mastropierro explores the interaction between corpus stylistics and translation studies. The book combines qualitative and quantitative methods to show how changes in the style of translation can affect the interpretation of a translated literary work compared to the original text.)

McIntyre, D. and Walker, B. (2019) Corpus Stylistics: Theory and Practice, Edinburgh: Edinburgh University Press. (This textbook provides an up-to-date introduction to corpus stylistics, a field that has seen many developments over the past decade. McIntyre and Walker guide the reader through hands-on examples for a range of registers, including fiction and non-fiction.)

Simpson, P. (2014) Stylistics: A Resource Book for Students, 2nd edn, London: Routledge. (The second edition of Simpson’s textbook includes new sections on recent trends in stylistics and has significantly expanded on corpus stylistics. This resource book contains units on foundational knowledge in the field, as well as ‘extension’ sections with reprints of publications by a range of scholars.)

Underwood, T. (2019) Distant Horizons: Digital Evidence and Literary Change, Chicago: University of Chicago Press. (Underwood’s monograph represents an approach to digital methods for analysing fiction that is rooted in literary culture and history. It introduces analyses of large amounts of data, larger than most corpus stylistic studies, and offers corpus linguists an opportunity to see connections between corpus research and distant reading.)


Corpora and the language of films: exploring dialogue in English and Italian

Berber Sardinha, T. and Veirano Pinto, M. (2019) Dimensions of Variation across American Television Registers, International Journal of Corpus Linguistics 24(1): 3–32. (Multi-dimensional analysis in this wide-ranging article is used to compare and group television programmes, including films, hence revealing the dimensions that appear to account for their variations.)


How to use corpus linguistics in sociolinguistics: a case study of modal verb use, age and change over time

Baker, P. (2010) Sociolinguistics and Corpus Linguistics, Edinburgh: Edinburgh University Press. (This book acts as a general primer for a range of ways that corpora can aid sociolinguistics, having chapters on demographic variation, comparing language use across different cultures and examining language change over time, studying transcripts of spoken interactions and identifying attitudes or discourses.)

Friginal, E. (ed.) (2017) Studies in Corpus-based Sociolinguistics, London: Routledge. (This edited collection of 14 chapters from a range of authors is divided into three sections: languages and dialects, social demographics and register characteristics.)


Murphy, B. (2010) Corpus and Sociolinguistics: Investigating Age and Gender in Female Talk, Amsterdam: John Benjamins. (This monograph involves a detailed analysis of age and gender in a 90,000-word spoken corpus of Irish English, considering features like hedges, vagueness, intensifiers and swearing.)


Corpus linguistics in the study of news media


Bednarek, M. and Caple, H. (2017) The Discourse of News Values. How Organizations Create Newsworthiness, Oxford: Oxford University Press. (This volume overcomes many of the limitations associated with corpus approaches to discourse, as well as to text-only approaches in general. The authors propose a multisemiotic approach to the study of media discourse and show how the newsworthiness of events is constructed discursively through words and images.)

Hansen, K. R. (2016) News from the Future: A Corpus Linguistic Analysis of Future-Oriented, Unreal and Counterfactual News Discourse, Discourse & Communication, 10(2): 115–136. (This study is ‘a grammatically founded approach to future-oriented journalism’ [p. 116], i.e. of journalism reporting on planned events and discussing consequences, expectations and agendas. By analysing patterns of use of modal verbs, verb tenses and speech acts, this longitudinal study of four Danish newspapers describes the shift towards a less event-centred and more analytical kind of journalism.)

Marchi, A. (2019) Self-Reflexive Journalism. A Corpus Study of Journalistic Culture and Community in the Guardian, London: Routledge. (This is a book-length study of how journalists define the boundaries of their community and set the standards of good vs. bad journalism in constructing and negotiating representations of their profession. The study has a strong methodological focus, employing and encouraging eclecticism and flexibility in the use of tools and constant scrutiny of the impact of research practices.)


How to use corpus linguistics in forensic linguistics
Coulthard, M., Johnson, A. and Wright, D. (2017) An Introduction to Forensic Linguistics: Language in Evidence, London: Routledge. (This book not only provides a comprehensive introduction to forensic linguistics in general, but it advocates for the use of corpus-based methods and gives a whole range of examples where it has been, and can be, applied.)
Cunningham, C. D. and Egbert, J. (2020) Using Empirical Data to Investigate the Original Meaning of “Emolument” in the Constitution, Georgia State University Law Review, 36(5): 465–489. (This paper provides an excellent overview of how corpus methods can be used to determine the meaning of a word and its implications for a real-life case.)
Heffer, C. (2005) The Language of Jury Trial: A Corpus-Aided Analysis of Legal-Lay Discourse, Basingstoke: Palgrave. (This book is perhaps the most comprehensive investigation into courtroom discourse that utilises corpus-based methods. Several different techniques are used to investigate how the various participants within a courtroom use language to their advantage.)
Larner, S. D. (2015) From Intellectual Challenges to Established Corpus Techniques: Introduction to the Special Issue on Forensic Linguistics, Corpora 10(2): 131–143. (This paper is an introduction to a special issue of Corpora on forensic linguistics. After a brief history of how the two fields came to work together, Larner presents an annotated bibliography of over 50 references, pointing towards work which meets at that intersection.)
Corpus linguistics in the study of political discourse: recent directions

The items chosen here are short article-length publications which the reader can use as case studies to explore methods (for book-length reports of projects, see Partington (2003) and Koteyko (2014)).

Baker, P. and McEnery, T. (2015) Who Benefits When Discourse Gets Democratised? Analysing a Twitter Corpus around the British Benefits Street Debate, in A. McEnery and P. Baker (eds) Corpora and Discourse Studies, London: Palgrave Macmillan, pp. 244–265. (This is an example of analysis which focusses on representation of a political issue in social media. The corpus is built from Twitter data, and the study uses keyness as the starting point for analysis.)

Clarke I. and Grieve J (2019) Stylistic Variation on the Donald Trump Twitter Account: A Linguistic Analysis of Tweets Posted between 2009 and 2018, PLoS ONE 14(9): e0222062 (This is an example of research which focusses on style in political discourse. The paper uses multidimensional analysis to track variation over time in Trump's tweets.)

Gabrielatos, C. (2020, July 28) Bibliography of Discourse-Oriented Corpus Studies, Retrieved from http://ehu.ac.uk/docsbiblio. (This useful online resource is continuously updated.)

Riihimäki, J. (2019) At the Heart and in the Margins: Discursive Construction of British National Identity in Relation to the EU in British Parliamentary Debates from 1973 to 2015, Discourse & Society 30(4), 412–431. (This is an example of diachronic research into parliamentary discourse. The starting point for the analysis is co-occurrence of lexical items.)


**Corpus linguistics and health communication: using corpora to examine the representation of health and illness**

Gwyn, R. (2002) Communicating Health and Illness, London: Sage. (This is an accessible discourse-based survey to health communication.)

Harvey, K. and Koteyko, N. (2012) Exploring Health Communication, London: Routledge. (This text provides an accessible overview of key concepts in language and health but also has more focus on corpus linguistic applications in this domain.)


Corpora in language testing: developments, challenges and opportunities

Barker, F. (2013) Using Corpora to Design Assessment, The Companion to Language Assessment 2: 1013–1028. (This article provides information on the use of corpora in large-scale test design.)

Cushing, S. T. (ed.) (2017) ‘Corpus Linguistics in Language Testing Research’ [Special issue], Language Testing 34(4). (This special issue presents five articles that use corpus linguistics tools and techniques to explore LTA issues, along with an introduction by the editor and commentaries by a corpus linguist and an LTA specialist.)

Park, K. (2014) Corpora and Language Assessment: The State of the Art, Language Assessment Quarterly 11(1): 27–44. (This article provides an overview of computational approaches to language assessment and advances in the use of corpora for LTA.)


Corpus linguistics and the study of social media: a case study using multi-dimensional analysis

Baker, P. and McEnery, T. (2015) Who Benefits When DiscourseGetsDemocratised? Analysing a TwitterCorpus around the British Benefits Street Debate, in P. Baker and T. McEnery (eds) Corpora and DiscourseStudies: Integrating Discourse and Corpora, Basingstoke: Palgrave Macmillan, pp. 244–265. (The authorsanalysed a corpus of tweets referring to the British TV show Benefits Street and to a televised debate aboutthe programme. The show centred on people receiving government support ["benefits"] who lived in a poorarea of Birmingham. The analysis detected some of the major discourses in the tweets, including "the idlepoor", which framed people in need as idle and undeserving. Overall, the study shows that social media corporaare valuable sources of data for corpus-based discourse studies.)

Clarke, I. and Grieve, J. (2017) Dimensions of Abusive Language on Twitter, Proceedings of the FirstWorkshop on Abusive Language Online, Vancouver, Canada, July 30 - August 4, 2017, pp. 1–10. (This paper presents an MD study of a corpus of 1,486 tweets coded for hate speech, such as racial, religious andsexist slurs. Through the MD analysis, the study shows that hate speech in social media is patterned forgrammar. In general, tweets displaying sexism are more interactive and attitudinal than tweets displayingracism.)

Rüdiger, S. and Dayter, D. (eds) (2020) Corpus Approaches to Social Media (Studies in Corpus Linguistics,Vol. 98), Amsterdam: John Benjamins. (This edited collection comprises papers dealing with severalimportant aspects of social media language, both verbal and visual. The volume includes case studies ofdifferent platforms such as Reddit, Twitter, WhatsApp and Facebook. The chapter by Christiansen, Danceand Wild tackles the challenges involved in analysing images, proposing the use of Google ArtificialIntelligence tools to carry out visual constituent analysis.)


Posthumanism and corpus linguistics


O'Halloran, K. A. (2017) Posthumanism and Deconstructing Arguments: Corpora and Digitally-Driven Critical Analysis, London: Routledge. (Provides an expanded account of the critical posthumanist reading strategy illustrated in Section 4, as well as of the strategy's relationship with critical discourse studies.)


