

Cambridge University Press  
978-1-107-03738-0 - The Cambridge Handbook of English Corpus Linguistics  
Edited by Douglas Biber and Randi Reppen  
Frontmatter  
[More information](#)

---

## The Cambridge Handbook of English Corpus Linguistics

*The Cambridge Handbook of English Corpus Linguistics (CHECL)* surveys the breadth of corpus-based linguistic research on English, including chapters on collocations, phraseology, grammatical variation, historical change, and the description of registers and dialects. The most innovative aspects of the *CHECL* are its emphasis on critical discussion, its explicit evaluation of the state of the art in each subdiscipline, and the inclusion of empirical case studies. While each chapter includes a broad survey of previous research, the primary focus is on a detailed description of the most important corpus-based studies in this area, with discussion of what those studies found, and why they are important. Each chapter also includes a critical discussion of the corpus-based methods employed for research in this area, as well as an explicit summary of new findings and discoveries.

DOUGLAS BIBER is Regents' Professor of Applied Linguistics in the English Department at Northern Arizona University.

RANDI REPPEN is Professor of Applied Linguistics in the English Department at Northern Arizona University.

Cambridge University Press

978-1-107-03738-0 - The Cambridge Handbook of English Corpus Linguistics

Edited by Douglas Biber and Randi Reppen

Frontmatter

[More information](#)

---

Cambridge University Press

978-1-107-03738-0 - The Cambridge Handbook of English Corpus Linguistics

Edited by Douglas Biber and Randi Reppen

Frontmatter

[More information](#)

# The Cambridge Handbook of English Corpus Linguistics

Edited by

**Douglas Biber**

and

**Randi Reppen**

*Northern Arizona University*



**CAMBRIDGE**  
UNIVERSITY PRESS

Cambridge University Press  
978-1-107-03738-0 - The Cambridge Handbook of English Corpus Linguistics  
Edited by Douglas Biber and Randi Reppen  
Frontmatter  
[More information](#)

**CAMBRIDGE**  
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

[www.cambridge.org](http://www.cambridge.org)

Information on this title: [www.cambridge.org/9781107037380](http://www.cambridge.org/9781107037380)

© Cambridge University Press 2015

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2015

Printing in the United Kingdom by TJ International Ltd. Padstow Cornwall

*A catalogue record for this publication is available from the British Library*

*Library of Congress Cataloguing in Publication data*

The Cambridge handbook of English corpus linguistics / edited by Douglas Biber and Randi Reppen  
Northern Arizona University.

p. cm. - (Cambridge handbooks in language and linguistics)

Includes bibliographical references and index.

ISBN 978-1-107-03738-0 (hardback)

1. Corpora (Linguistics) 2. Linguistic analysis (Linguistics) 3. Computational linguistics. I. Biber, Douglas, editor. II. Reppen, Randi, editor.

P128.C68C46 2015

420.1'88-dc23

2014035299

ISBN 978-1-107-03738-0 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Cambridge University Press  
978-1-107-03738-0 - The Cambridge Handbook of English Corpus Linguistics  
Edited by Douglas Biber and Randi Reppen  
Frontmatter  
[More information](#)

---

*We dedicate this handbook to the memory of Geoffrey Leech, one of the true pioneers of English corpus linguistics. Geoff contributed seminal research to all aspects of corpus linguistics, from corpus construction to corpus tagging and annotation, to the evaluation of corpus representativeness. But from our point of view, Geoff's most important contributions involved the application of corpus-based analysis to investigate linguistic research questions about language variation and use. Geoff contributed to nearly all major subdisciplines of English linguistics with such analyses, including pragmatics, stylistics, descriptive grammar, dialect differences, and historical change. For us personally, Geoff was a model, mentor, co-author, and friend. He will be sorely missed.*

Cambridge University Press

978-1-107-03738-0 - The Cambridge Handbook of English Corpus Linguistics

Edited by Douglas Biber and Randi Reppen

Frontmatter

[More information](#)

---

# Contents

<i>List of figures</i>	page ix
<i>List of tables</i>	xii
<i>Major corpora cited in the handbook</i>	xv
<i>List of contributors</i>	xvii
Introduction <i>Douglas Biber and Randi Reppen</i>	1
<b>Part I</b> Methodological considerations	9
1 Corpora: an introduction <i>Mark Davies</i>	11
2 Computational tools and methods for corpus compilation and analysis <i>Paul Rayson</i>	32
3 Quantitative designs and statistical techniques <i>Stefan Th. Gries</i>	50
<b>Part II</b> Corpus analysis of linguistic characteristics	73
4 Discourse intonation: a corpus-driven study of prominence on pronouns <i>Winnie Cheng</i>	75
5 Keywords <i>Jonathan Culpeper and Jane Demmen</i>	90
6 Collocation <i>Richard Xiao</i>	106
7 Phraseology <i>Bethany Gray and Douglas Biber</i>	125
8 Descriptive grammar <i>Geoffrey Leech</i>	146
9 Grammatical variation <i>Daniela Kolbe-Hanna and Benedikt Szmrecsanyi</i>	161
10 Grammatical change <i>Martin Hilpert and Christian Mair</i>	180
11 Lexical grammar <i>Susan Hunston</i>	201
12 Using corpora in discourse analysis <i>Alan Partington and Anna Marchi</i>	216
13 Pragmatics <i>Brian Clancy and Anne O'Keeffe</i>	235
14 Historical pragmatics <i>Irma Taavitsainen</i>	252

---

<b>Part III Corpus analysis of varieties</b>	269
15 Spoken discourse <i>Shelley Staples</i>	271
16 Corpora and written academic English <i>Ken Hyland</i>	292
17 Register variation <i>Susan Conrad</i>	309
18 Diachronic registers <i>Merja Kytö and Erik Smitterberg</i>	330
19 Literary style and literary texts <i>Michaela Mahlberg</i>	346
20 Dialect variation <i>Jack Grieve</i>	362
21 World Englishes <i>Marianne Hundt</i>	381
22 New answers to familiar questions: English as a lingua franca <i>Anna Mauranen, Ray Carey, and Elina Ranta</i>	401
23 Learner language <i>Gaëtanelle Gilquin and Sylviane Granger</i>	418
<b>Part IV Other applications of corpus analysis</b>	437
24 Vocabulary <i>Ron Martinez and Norbert Schmitt</i>	439
25 Lexicography and phraseology <i>Magali Paquot</i>	460
26 Classroom applications of corpus analysis <i>Thomas Cobb and Alex Boulton</i>	478
27 Corpus versus non-corpus-informed pedagogical materials: grammar as the focus <i>Fanny Meunier and Randi Reppen</i>	498
28 Translation <i>Silvia Bernardini</i>	515
<i>References</i>	537
<i>Index</i>	618



# Figures

1.1. Decrease in <i>for</i> (as conjunction in COHA), 1810s–2000s	17
1.2. Decrease in <i>for</i> (as conjunction in COCA), 1990s–2012	17
1.3. “Snippet” view in Google (Web)	20
1.4. Frequency chart in Google Books	21
1.5. Lexical frequency: Sketch Engine	23
1.6. Lexical frequency: Google Books (BYU)	23
1.7. Word forms ( <i>*ism</i> ) in Google Books (BYU)	23
1.8. Morphological variation in Google Books (BYU)	23
1.9. Probabilistic POS tagging in Google Books (BYU)	24
1.10. Syntactic searches in Sketch Engine	25
1.11. Syntactic searches in Google Books (BYU)	25
1.12. Lexis of fiction and newspapers in COCA	26
1.13. <i>*al.[j*]</i>	26
1.14. <i>[vv*] to</i>	26
1.15. <i>get</i> passive	27
1.16. quotative <i>like</i> : <i>[c*] [p*] [be] like,</i>	27
1.17. <i>[be] real [j*] [y*]</i>	27
1.18. Collocates of <i>chair</i> in fiction and newspapers in COCA	27
2.1. Concordance example for the word <i>support</i>	40
3.1. Hypothetical and actual frequencies of the forms of <i>GIVE</i> in the ICE-GB and their relative entropies ( $H_{rel}$ )	52
3.2. The effect of <i>VARIETY ON TENSE</i>	64
3.3. Cluster-analytic results for English consonant phonemes (from Gries 2013a)	69
4.1. Sample concordance lines for <i>most + people</i>	87
4.2. Sample concordance lines for <i>just + only</i>	88
5.1. A keyword dispersion plot of <i>Romeo and Juliet</i> generated by <i>WordSmith Tools</i> (reproduced from Scott and Tribble 2006: 65)	93
7.1. Number of distinct four-word combinations occurring at least once (with no dispersion restriction)	140

7.2. Number of distinct four-word pattern types occurring at two frequency levels (range = 5 texts)	141
7.3. Relationship between variability and predictability for 12*4 frames occurring more than 200 times per million words in academic prose	143
9.1. Projection of intercept adjustments to geography	177
10.1. Possible results for analyses of the Brown family of corpora	183
10.2. The decline of the modal auxiliaries (based on Leech 2003: 228, table 3)	186
10.3. Absolute and normalized frequencies of new types with <i>-ment</i> in the <i>OED</i>	192
10.4. Changes in the expanding productivity of <i>-ment</i> in the <i>OED</i> (Hilpert 2013: 131)	194
12.1. How the briefings participants refer to the Libyan administration in the first three months of 2011	228
13.1. Concordance lines for <i>mam</i> in <i>SettCorp</i> (sorted 2R then 3R)	246
13.2. Functions of vocatives in <i>Liveline</i> and <i>SettCorp</i> (normalized to 10,000 words)	248
14.1. General references in the text categories of EMENT 1500–1700 (according to Taavitsainen 2009)	266
15.1. Overall trends for stance features across phases and speaker groups	286
15.2. Frequency of modals used across phases and speaker groups	286
15.3. Frequency of stance adverbs used across phases and speaker groups	287
15.4. Frequency of stance complement clauses used across phases and speaker groups	290
17.1. Linguistic feature and text category emphases in register variation studies	311
17.2. Register comparison along two dimensions Dimension 1: involved vs. informational production Dimension 5: abstract/impersonal vs. non-impersonal style	325
20.1. Temporal variation in <i>not</i> contraction	374
20.2. Gender variation in <i>not</i> contraction	375
20.3. <i>do not</i> contraction	375
20.4. <i>Not</i> contraction local spatial autocorrelation maps	377
20.5. <i>Not</i> contraction clusters	378
20.6. Regional variation in <i>not</i> contraction	379
21.1. Categorization of PP contexts (Davydova 2011: 124)	389
21.2a. Relative frequency of PP and SP in the press section of ICE corpora (parsed)	394
21.2b. PPs (frequency pmw) in the press sections of ICE corpora (parsed)	395
21.3a. PP vs. SP with selected verbs	396
21.3b. PP vs. SP with selected verbs (active only)	397

23.1. Relative frequency per 100,000 words of <i>sort of</i> in the different subcorpora	432
23.2. Relative frequency per 100,000 words of <i>in fact</i> in the different subcorpora	433
24.1. Entry for <i>game</i> in the West (1953) General Service List	441
24.2. A sample of unedited 2–4 grams list derived from BNC	453
24.3. Example of initial data deletion phase (faded <i>n</i> -grams are deleted ones)	453
24.4. A sample from the first draft of the PHRASE List	455
24.5. A sample from a later revision of the PHRASE List	456
24.6. Example of integrated list of phrasal expressions and single words	456
24.7. Sample of the PHRASE List with numerical genre-sensitive frequency information	457
24.8. Genre-sensitive frequency information represented by system of symbols	458
25.1. A sample of the WordSketch for the verb <i>suggest</i>	462
28.1. Box plots for <b>Adverb-Verb</b> (MI and LOG FQ)	532
28.2. Box plot for <b>Adverb-Adjective</b> (MI)	532
28.3. Box plot for <b>Adjective-Noun</b> (LOG FQ)	533
28.4. Box plot for <b>Noun-(Preposition OR Conjunction)-Noun</b> (LOG FQ)	533
28.5. Box plot for <b>Adverb-Verb</b> (MI)	534

# Tables

1.1. Types of phenomena	13
1.2. Frequency of different phenomena in COCA, BNC, and Brown	14
1.3. Phenomena that can be researched with three text archives / Web	19
1.4. Frequency of very infrequent words in BNC, COCA, and three text archives / Web	20
1.5. Phenomena that can be researched with two “hybrid” corpora	22
1.6. Number of collocates in different corpora	25
1.7. Similarity of lexis in web-based GloWbE and genres in COCA and BNC	28
2.1. Quantitative analysis of papers published in 2012	35
2.2. Top 10 <i>n</i> -grams from <i>Alice’s Adventures in Wonderland</i>	45
2.3. C-gram tree view for <i>and the</i>	47
3.1. The frequencies of several <i>n</i> -grams in the untagged Brown corpus	54
3.2. Damerou’s (1993) relative frequency ratio	55
3.3. Schematic co-occurrence table of token frequencies for association measures	56
3.4. Toy example for 3L-3R collocations of <i>the</i> with row and column entropies	57
3.5. Co-occurrence table for <i>of</i> and <i>course</i> in the spoken component of the BNC	58
3.6. The distribution of different types of NPs across subject/non-subject slots (Aarts 1971: table 4.5)	61
3.7. Hundt and Smith’s (2009) observed frequencies of English present perfects and simple pasts in LOB, FLOB, Brown, and Frown	63
3.8. Dimensions of variation in Biber (1988)	68
3.9. Chi-squared values with 2 <i>df</i> for pairwise comparisons (Egan 2012: table 1)	70

4.1. Personal pronouns in HKCSE (prosodic)	85
4.2. Possessive pronouns in HKCSE (prosodic)	86
5.1. Romeo's parts-of-speech rank-ordered for positive keyness (i.e. relative overuse)	102
5.2. Romeo's semantic categories rank-ordered for positive keyness (i.e. relative overuse)	104
6.1. Contingency table	111
6.2. Examples of semantic prosodies	113
6.3. Distribution of <i>CONSEQUENCE</i> across meaning categories in FLOB/Frown	119
7.1. Major design parameters of corpus-based and corpus-driven studies of phraseology	126
7.2. Research studies of extended lexical sequences	134
7.3. Summary of lexical bundles and frame patterns investigated in the case study	139
7.4. 12*4 Frames identified in the present study in academic writing	142
9.1. Logistic regression model with fixed and random effects in complementation choice. Predicted odds are for the retention of the <i>that</i> -complementizer	175
10.1. Periods identified through VNC	195
10.2. Input data for a HCFA	197
10.3. Results for types of <i>-ment</i> formation	198
13.1. Occurrences of vocatives in <i>Liveline</i> and <i>SettCorp</i> , normalized to 10,000 words	247
15.1. Composition of the corpus used for the study	284
15.2. Lexico-grammatical features used for stance analyses	284
15.3. Phases of the interactions	285
16.1. Selected features in research articles and textbooks	293
17.1. Examples of studies focusing on individual features: single register, and comparisons of registers and subregisters	312
17.2. Examples of studies using multidimensional (MD) analysis	317
17.3. Examples of ESP studies that compare students and proficient speakers/writers	319
17.4. Registers in civil engineering used in the study	323
17.5. Features on Biber's (1988) multidimensional analysis of English (with factor loadings)	324
18.1. <i>Thou</i> and <i>you</i> forms in trials, depositions and drama 1560–1760 (Walker 2007): raw frequencies and row percentages	342
20.1. Gender variation in <i>not</i> contraction	374
20.2. Regional variation in <i>not</i> contraction	378
21.1. PP vs. SP in present perfect contexts	390

21.2. Forms expressing perfect meaning	390
21.3. Co-occurrence of PP and SP with adverbials <i>just, (n)ever, yet</i>	397
23.1. Relative frequency per 100,000 words of DMs in LOCNEC and LINDSEI	432
24.1. Comparison of new GSL to West GSL and AWL (Brezina and Gablasova 2013)	444
24.2. Coverage of AVL and AWL in COCA academic and BNC academic (Gardner and Davies 2013: 19)	445
24.3. Spoken AFL top 10	450
24.4. 1,000-level frequency cut-offs (BNC)	452
25.1. Collocation boxes for verbs of evidence in the Big Five	471
25.2. Recall rates for the collocates of the verb <i>support</i> in <i>LDOCE5</i> , <i>CCAD6</i> , and <i>MEDAL2</i>	473
25.3. Precision rates for subject and object collocations of verbs of evidence in <i>LDOCE5</i> , <i>CCAD6</i> , and <i>MEDAL2</i>	474
26.1. Within-groups effect size ( $k = 8$ ), sorted by effect size	492
26.2. Between-groups effect size ( $k = 13$ ), sorted by effect size	493
27.1. The passive in four corpus-informed grammar books	507
27.2. The passive in four non-corpus-informed grammar books	508
27.3. Comparison of the responses per category in Tables 27.1 and 27.2	509
27.4. Exercises on the passive in the four corpus-informed books	511
28.1. The <i>FINREP</i> corpus	527
28.2. The <i>SHARLET</i> corpus	528
28.3. POS patterns used for candidate collocation extraction	529
28.4. Adverb-Verb rankings (bigrams from translated and non-translated English financial reports, with frequency and MI data from <i>ukWaC</i> )	530
28.5. Results of significance testing (comparison of rankings from the <i>FINREP-NT-EN</i> and <i>FINREP-TR-EN</i> subcorpora)	531
28.6. Results of significance testing (comparison of rankings from the <i>SHARLET-NT-EN</i> and <i>SHARLET-TR-EN</i> subcorpora)	534

# Major corpora cited in the handbook

ARCHER	<i>A Representative Corpus of Historical English Registers</i>
BNC	<i>British National Corpus</i>
Brown	<i>Brown Corpus</i>
CED	<i>Corpus of English Dialogues</i>
CEEC	<i>Corpus of Early English Correspondence</i>
CIC	<i>Cambridge International Corpus</i>
COCA	<i>Corpus of Contemporary American English</i>
COHA	<i>Corpus of Historical American English</i>
COLT	<i>Bergen Corpus of London Teenager Language</i>
ELFA	<i>Corpus of English as a Lingua Franca in Academic Settings</i>
EMEMT	<i>Early Modern English Medical Texts</i>
FLOB	<i>Freiburg-LOB Corpus</i>
FRED	<i>Freiburg English Dialect Corpus</i>
FROWN	<i>Freiburg-Brown Corpus</i>
GloWbE	<i>Corpus of Global Web-based English</i>
HKCSE	<i>Hong Kong Corpus of Spoken English</i>
ICE	<i>International Corpus of English</i>
ICLE	<i>International Corpus of Learner English</i>
LINDSEI	<i>Louvain International Database of Spoken English Interlanguage</i>
LLC	<i>London-Lund Corpus</i>
LOB	<i>Lancaster-Oslo-Bergen Corpus</i>
LOCNEC	<i>Louvain Corpus of Native English Conversation</i>
LSWC	<i>Longman Spoken and Written English Corpus</i>
LWP	<i>Language in the Workplace Project</i>
MICASE	<i>Michigan Corpus of Spoken Academic English</i>
MICUSP	<i>Michigan Corpus of Upper-level Student Papers</i>

NHCC	<i>Nottingham Health Communication Corpus</i>
T2KSWAL	<i>TOEFL 2000 Spoken and Written Academic Language Corpus</i>
Time	<i>Time Magazine Corpus of American English</i>
VOICE	<i>Vienna–Oxford International Corpus of English</i>
W <sub>r</sub> ELFA	<i>The Written Corpus of English as a Lingua Franca in Academic Settings</i>



# Contributors

Silvia Bernardini, Associate Professor of English Language and Translation,  
University of Bologna

Douglas Biber, Regents' Professor, Applied Linguistics, Northern Arizona  
University

Alex Boulton, Professor, English and Applied Linguistics,  
Université de Lorraine

Ray Carey, Researcher, ELFA Project, University of Helsinki

Winnie Cheng, Professor, English language, Hong Kong Polytechnic  
University

Brian Clancy, Researcher, Mary Immaculate College, University of  
Limerick

Thomas Cobb, Associate Professor, Applied Linguistics, Université du  
Québec à Montréal, Canada

Susan Conrad, Professor, Applied Linguistics, Portland State University

Jonathan Culpeper, Professor, Linguistics and English Language, Lancaster  
University

Mark Davies, Professor, Linguistics, Brigham Young University

Jane Demmen, Research Fellow, Linguistics and Modern Languages,  
University of Huddersfield

Gaëtanelle Gilquin, Research Associate, FNRS, Centre for English Corpus  
Linguistics, Université catholique de Louvain

Sylviane Granger, Professor, Centre for English Corpus Linguistics,  
Université catholique de Louvain

Bethany Gray, Assistant Professor, Applied Linguistics and Technology,  
Iowa State University

Stefan Th. Gries, Professor, Linguistics, University of California, Santa  
Barbara

Jack Grieve, Lecturer, Forensic Linguistics, Aston University

Martin Hilpert, Assistant Professor, English Linguistics, University of  
Neuchâtel

- 
- Marianne Hundt, Professor and Chair, English Language and Linguistics,  
 University of Zurich  
 Susan Hunston, Professor, English Language and Applied Linguistics,  
 University of Birmingham  
 Ken Hyland, Chair, Professor of Applied Linguistics and Director of the  
 Center for Applied English Studies, University of Hong Kong  
 Daniela Kolbe-Hanna, Assistant Professor, English Studies, University of  
 Trier  
 Merja Kytö, Professor, English Language, Uppsala University  
 Geoffrey Leech, Professor Emeritus, Linguistics and English Language,  
 Lancaster University  
 Michaela Mahlberg, Professor, English Language and Linguistics,  
 University of Nottingham  
 Christian Mair, Professor and Chair, English Linguistics, University of  
 Freiburg  
 Anna Marchi, Linguistics and English Language, Lancaster University  
 Ron Martinez, Assistant Professor of English, Federal University of Paraná  
 Anna Mauranen, Professor of English Philology and Vice Rector, University  
 of Helsinki  
 Fanny Meunier, Professor, Centre for English Corpus Linguistics,  
 Université catholique de Louvain,  
 Anne O'Keefe, Senior Lecturer, Applied Linguistics and English Language  
 Teaching, Mary Immaculate College, University of Limerick  
 Magali Paquot, Postdoctoral Researcher, FNRS, Centre for English Corpus  
 Linguistics, Université catholique de Louvain  
 Alan Partington, Associate Professor, Linguistics, University of Bologna  
 Elina Ranta, Researcher, ELFA Project University of Helsinki, University of  
 Tampere  
 Paul Rayson, Director, University Centre for Computer Corpus Research on  
 Language (UCREL), Lancaster University  
 Randi Reppen, Professor, Applied Linguistics and TESL, Northern Arizona  
 University  
 Norbert Schmitt, Professor, Applied Linguistics, University of Nottingham  
 Erik Smitterberg, Researcher, English Language, Uppsala University  
 Shelley Staples, Assistant Professor, English, Purdue University  
 Benedikt Szmrecsanyi, Odysseus Research Professor, Linguistics, KU  
 Leuven  
 Irma Taavitsainen, Professor, English Philology, University of Helsinki  
 Richard Xiao, Reader, Linguistics and English Language, Lancaster  
 University