Corpora in Applied Linguistics

Corpus linguistics has revolutionised the world of language study and is an essential component of work in applied linguistics. This book, now in its second edition, provides a thorough introduction to all the key research issues in corpus linguistics, from the point of view of applied linguistics. The field has progressed a great deal since the first edition, so this edition has been completely rewritten to reflect these advances, whilst still maintaining the emphasis on hands-on corpus research of the first edition. It includes chapters on qualitative and quantitative research, applications in language teaching, discourse studies, and beyond. It also includes an extensive discussion of the place of corpus linguistics in linguistic theory, and provides numerous detailed examples of corpus studies throughout. Providing an accessible but thorough grounding to the fascinating, fast-moving field of corpus linguistics, this book is essential reading for the student and the researcher alike.

Susan Hunston is Professor of English Language at the University of Birmingham, UK. Her recent publications include *Corpus Approaches to Evaluation* (Routledge, 2011) and *Interdisciplinary Research Discourse* (with Paul Thompson, Routledge, 2019). She is a former Chair of the British Association for Applied Linguistics.

THE CAMBRIDGE APPLIED LINGUISTICS SERIES

The authority on cutting-edge Applied Linguistics research

Series Editors 2007–present: Carol A. Chapelle and Susan Hunston
1988–2007: Michael H. Long and Jack C. Richards

For a complete list of titles please visit: www.cambridge.org

Recent titles in this series

Corpora in Applied Linguistics
2nd edition
Susan Hunston

The Language of Mental Illness
Corpus Linguistics and the Construction of Mental Illness in the Press
Hazel Price

Mobile Assisted Language Learning
Concepts, Contexts and Challenges
Glenn Stockwell

Research Genres Across Languages
Multilingual Communication Online
Carmen Pérez-Llantada

Validity Argument in Language Testing
Case Studies of Validation Research
Edited by Carol A. Chapelle and Erik Voss

Doing English Grammar
Theory, Description and Practice
Roger Berry

Learner Corpus Research Meets Second Language Acquisition
Bert Le Bruyn and Magali Paquot

Second Language Speech Fluency
From Research to Practice
Parvaneh Tavakoli and Clare Wright

Ontologies of English
Conceptualising the Language for Learning, Teaching, and Assessment
Edited by Christopher J. Hall and Rachel Wicaksono

Task-Based Language Teaching
Theory and Practice
Rod Ellis, Peter Skehan, Shaofeng Li, Natsuko Shintani and Craig Lambert

Feedback in Second Language Writing
Contexts and Issues
2nd edition
Edited by Ken Hyland and Fiona Hyland

Language and Television Series
A Linguistic Approach to TV Dialogue
Monika Bednarek

Intelligibility, Oral Communication, and the Teaching of Pronunciation
John M. Levis

Multilingual Education
Between Language Learning and Translanguaging
Edited by Cenoz Jasone and Gorter Durk

Learning Vocabulary in Another Language
2nd edition
I. S. P. Nation

Narrative Research in Applied Linguistics
Edited by Barkhuizen Gary

Teacher Research in Language Teaching
A Critical Analysis
Simon Borg

Figurative Language, Genre and Register
Alice Deignan, Jeannette Littlemore and Elena Semino

Exploring ELF
Academic English Shaped by Non-native Speakers
Anna Mauranen

Genres across the Disciplines
Student Writing in Higher Education
Hilary Nesi and Sheena Gardner

Disciplinary Identities
Individuality and Community in Academic Discourse
Ken Hyland

Replication Research in Applied Linguistics
Edited by PorteGraeme

The Language of Business Meetings
Michael Handford

Reading in a Second Language
Moving from Theory to Practice
William Grabe

Modelling and Assessing Vocabulary Knowledge
Edited by Daller Helmut, Milton James and Treffers-Daller Jeanine

Practice in a Second Language
Perspectives from Applied Linguistics and Cognitive Psychology
Edited by DeKeyser Robert M.

Task-Based Language Education
From Theory to Practice
Edited by van den Branden Kris

Second Language Needs Analysis
Edited by Long Michael H.

Insights into Second Language Reading
A Cross-Linguistic Approach
Keiko Koda

Research Genres
Exploration and Applications*John M. Swales*

Critical Pedagogies and Language Learning
Edited by Norton Bonny and Toohey Kelleen

Exploring the Dynamics of Second Language Writing
Edited by Kroll Barbara

Understanding Expertise in Teaching
Case Studies of Second Language Teachers
Amy B. M. Tsui

Criterion-Referenced Language Testing
James Dean Brown and Thom Hudson

Corpora in Applied Linguistics
Susan Hunston

Pragmatics in Language Teaching
Edited by Rose Kenneth R. and Kasper Gabriele

Cognition and Second Language Instruction
Edited by Robinson Peter

Research Perspectives on English for Academic Purposes
Edited by Flowerdew John and Peacock Matthew

Computer Applications in Second Language Acquisition
Foundations for Teaching, Testing and Research
Carol A. Chapelle

Cambridge University Press & Assessment
978-1-108-42509-4 — Corpora in Applied Linguistics
Susan Hunston
Frontmatter
More Information

Corpora In Applied Linguistics

Susan Hunston
University of Birmingham

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

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
Information on this title: www.cambridge.org/9781108425094
DOI: 10.1017/9781108616218

© Susan Hunston 2022

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 2022

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

Library of Congress Cataloging-in-Publication Data
Names: Hunston, Susan, 1953– author.
Title: Corpora in applied linguistics / Susan Hunston.
Description: Second edition. | Cambridge, United Kingdom ; New York, NY :
Cambridge University Press, 2022. | Series: The Cambridge applied linguistics series |
Includes bibliographical references and index.
Identifiers: LCCN 2022001489 (print) | LCCN 2022001490 (ebook) |
ISBN 9781108425094 (hardback) | ISBN 9781108441186 (paperback) |
ISBN 9781108616218 (epub)
Subjects: LCSH: Corpora (Linguistics) | Applied linguistics. | BISAC: LANGUAGE
ARTS & DISCIPLINES / Linguistics / General
Classification: LCC P128.C68 H85 2022 (print) | LCC P128.C68 (ebook) |
DDC 410.1/88–dc23/eng/20220208
LC record available at https://lccn.loc.gov/2022001489
LC ebook record available at https://lccn.loc.gov/2022001490

ISBN 978-1-108-42509-4 Hardback
ISBN 978-1-108-44118-6 Paperback

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.

For Michael Hoey, 1948–2021

Cambridge University Press & Assessment
978-1-108-42509-4 — Corpora in Applied Linguistics
Susan Hunston
Frontmatter
More Information

Contents

List of Figures xii
List of Tables xiv
Acknowledgements xviii

1 *Introduction* 1
1.1 Corpus Linguistics 1
1.2 About this Book 3
1.3 Terminology 5
1.4 Commonly Used Resources 9
1.5 Corpus Linguistics: A Personal View 10

2 *Designing a Corpus* 18
2.1 A Historical Perspective on Corpus Design 18
2.2 Types of Corpus 21
2.3 Issues in Designing or Selecting a Corpus 30
2.4 Corpus Source Material and Issues 38
2.5 Conclusion 43
Appendix: List of Some Commonly Used Corpora of English 44

3 *Learning from a Corpus: Finding Pattern in Concordance Lines* 47
3.1 Introduction 47
3.2 Phraseology and Meaning in Concordance Lines 51
3.3 'Hidden Meaning' in Concordance Lines 55
3.4 Types of Patterning 60
3.5 Conclusion 79

4 *Foundational Quantitative Concepts in Corpus Linguistics* 81
4.1 Introduction 81
4.2 Frequency, Relative Frequency and Comparisons 82
4.3 Frequency and Normalisation 84
4.4 Keywords 88
4.5 Measuring Collocation 93

© in this web service Cambridge University Press & Assessment www.cambridge.org

4.6 Lexical Bundles 102
4.7 Multidimensional Analysis 105
4.8 Semantic and Other Annotation 111
4.9 Conclusion 114

5 *Exploring Corpora through Quantitative Measures* 116
5.1 Introduction 116
5.2 More about Keyness 117
5.3 Collostructions 121
5.4 Clustering Texts 126
5.5 Clustering Words 135
5.6 Conclusion 140

6 *The Language Learner and the Corpus* 143
6.1 Introduction 143
6.2 Reference Materials for Learners: Dictionaries and Grammars 143
6.3 Learner Corpora 152
6.4 Language Learners Using Corpora 174
6.5 Conclusion 184

7 *Corpora and the Analysis of Discourse* 186
7.1 Introduction 186
7.2 Academic English: Comparing Disciplines 187
7.3 Corpus Linguistics and News Discourse: Being Critical 210
7.4 Stylistics 217
7.5 Further Applications of Corpus Studies of Discourse 221
7.6 Conclusion 224

8 *Further Applications of Corpus Research* 226
8.1 Introduction 226
8.2 Applications of Bilingual Corpora 226
8.3 Corpora in Forensic Linguistics 233
8.4 Information and Opinion Extraction 241
8.5 Identifying Fake News 245
8.6 Sociolinguistics: The Study of Variation in Language 247
8.7 Conclusion 253

9 *Corpora and Language Theory* 255
9.1 Introduction 255
9.2 Sinclair and Language Theory 256
9.3 The Mental Modelling of Language 258

© in this web service Cambridge University Press & Assessment

Cambridge University Press & Assessment
978-1-108-42509-4 — Corpora in Applied Linguistics
Susan Hunston
Frontmatter
More Information

9.4 Language as Social Semiotic 268
9.5 Conclusion 286

10 *Conclusion* 288
10.1 First Catch Your Corpus 289
10.2 Investigation Methods: Top-down 292
10.3 Investigation Methods: Bottom-up 295
10.4 And the Outcome Is? 303
10.5 Conclusion 304

References 307
Index 334

© in this web service Cambridge University Press & Assessment www.cambridge.org

Figures

1.1 Concordance lines for *possible* from Wordbanks Online 11
3.1a Twenty instances of *discussion/s* from the BNC (original order) 49
3.1b Twenty instances of *discussion/s* from the BNC (right-sorted) 49
3.1c Twenty instances of *discussion/s* from the BNC (left-sorted) 49
3.2a Concordance lines for *initiative* sense 1 (Times newspapers in BoE) 52
3.2b Concordance lines for *initiative* sense 2 (Economist magazine in BoE) 52
3.2c Concordance lines for *initiative* sense 3 (Spoken English in BoE) 52
3.3a Concordance lines for condemn from the BoE, meaning 'criticise' 53
3.3b Instances of condemn from the BoE, meaning 'pass sentence' 53
3.3c Instances of condemn from the BoE, meaning 'having been sentenced to death' 53
3.3d Instances of condemn from the BoE, meaning 'bring about unfortunate circumstances' 54
3.4 Concordance lines for 'it be adjective that' from the BNC 55
3.5 Concordance lines for *move* from the BNC 57
3.6 Concordance lines for *budge* from the BNC 57
3.7a Concordance lines for *have one's cake and eat it* from the BNC 57
3.7b Lines from 3.7a with negative modality 58
3.7c Lines from 3.7a with WANT or TRY 58
3.7d Counter-examples from 3.7a 58
3.8 Patterns of use of the noun *discrepancy* (English Web 2013 sample) 61
3.9 Concordance lines for *struggle* (BoE) right-sorted 66
3.10 Concordance lines for *the argument that* from Birmingham-Elsevier Environment Corpus 76

© in this web service Cambridge University Press & Assessment www.cambridge.org

Cambridge University Press & Assessment
978-1-108-42509-4 — Corpora in Applied Linguistics
Susan Hunston
Frontmatter
More Information

4.1 Concordance lines for *species* from *The Rough Guide to Evolution*, with ±4 words in bold 94
4.2 Text types shown on the 'Involved versus Informational Production' dimension. Adapted from 'Variation across Speech and Writing', Biber 1988. Reproduced with permission of the Licensor through PLSclear 109
4.3 Text types shown on the 'Narrative versus Non-narrative Concerns' dimension. Adapted from 'Variation across Speech and Writing', Biber 1988. Reproduced with permission of the Licensor through PLSclear 110
5.1 VIS map representing degrees of overlap between authors. Adapted from 'Lexical bundles and disciplinary variation in university students' writing' (Durrant 2017) Reproduced with permission of the Licensor through Rightslink 128
5.2 A sample dendrogram 132
5.3 Dimension profiles of constellations. Adapted from 'Multi-Dimensional analysis, text constellations, and interdisciplinary discourse' (Thompson et al. 2017). Reproduced with permission of John Benjamins 133
6.1 The noun *controversies* followed by *about* and *over* 176
8.1 All uses of *then* in Bentley's statement 234
9.1 A simplified mood network. Adapted from 'Introduction to Function Grammar' (Halliday and Matthiessen 2014). Reproduced with permission of the Licensor through PLSclear 269
9.2 Suggested Attitude network. Adapted from 'Language patterns and attitude revisited' (Su and Hunston 2019b). Reproduced with permission of John Benjamins 277
9.3 Network of process types. Adapted from 'Introduction to Function Grammar' (Halliday and Matthiessen 2014). Reproduced with permission of the Licensor through PLSclear 278
9.4 A verbal process network based on pattern alternatives 278
9.5 A pattern-based network for ASK and ENQUIRE 279
9.6 A network for the EXCHANGE meaning 283
9.7 Mid-point of the network based on V n *with* n 285
9.8 End-point of the network based on V n *with* n 286
10.1 Selected concordance lines illustrating 'common effort' from the Coronavirus Corpus 299
10.2 Selected concordance lines for COMBAT *vaccine hesitancy* from the Coronavirus Corpus, April 2021 303

© in this web service Cambridge University Press & Assessment www.cambridge.org

Tables

1.1 Ngrams in a sentence 8
1.2 Patterns of *possible* in 250 random lines from Wordbanks Online 13
1.3 The adjective *possible* compared with other adjectives 16
2.1 A parallel corpus with a single source text 27
2.2 A parallel corpus of original and translated texts 27
3.1 Tabular format of selected examples from Figure 3.1 50
3.2 Patterns in Example 3.4 64
3.3 Hierarchical analysis of Example 3.4 65
3.4 Sequential analysis of Example 3.4 65
3.5 *struggle*: pattern and meaning 67
3.6 Examples and analyses of the REVENGE Frame, adapted from the FrameNet website 70
3.7 PDEV coding for the verb ENCOURAGE, adapted from pdev.org 71
3.8 Meaning mapped on to pattern: **V n *of* n**. Adapted from ‘Patterns, constructions, and applied linguistics’ (Hunston 2019). Reproduced with permission of John Benjamins. 73
3.9 Part of a local grammar of evaluation. Adapted from ‘Patterns, constructions, and local grammar’ (Hunston and Su 2019). Reproduced with permission of the Licensor through Rightslink. 75
4.1 Percentage use of *appearance* and *disappearance* in two corpora 84
4.2 Raw and normalised lemma frequencies in two corpora 85
4.3 Nouns with a normalised frequency of 10–11, showing Range and Dispersion (Leech et al. 2001) 87
4.4 The most frequent lexical collocates of *species* in *The Rough Guide to Evolution* 95
4.5 Most significant 20 collocates of *species* in *The Rough Guide to Evolution* using three measures (from AntConc) 96

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE

Cambridge University Press & Assessment
978-1-108-42509-4 — Corpora in Applied Linguistics
Susan Hunston
Frontmatter
More Information

4.6 Examples of collocation pairs for EAP. Adapted from 'The academic English collocation list: a corpus-driven study' (Lei and Liu 2018) 99
4.7 Sample from Word Sketch for *pollution* in environment journals (from SketchEngine) 100
4.8 Sample categories from Table 4.7 showing most significant and most frequent 101
4.9 The 4-grams from *Global Environmental Change* with a frequency over 160 (40pmw) 104
4.10 Two factors identified through multidimensional analysis. Adapted from 'Variation across Speech and Writing' (Biber 1988). Reproduced with permission of the Licensor through PLSclear. 107
5.1 Ratio of hypothetical words in two hypothetical corpora 119
5.2 Imagined word frequencies and text use in two hypothetical corpora 120
5.3 Co-occurrence table. Adapted from '15 years of collostructions' (Gries 2019). Reproduced with permission of John Benjamins. 122
5.4 Example of a populated co-occurrence table 122
5.5 Example collostructions 123
5.6 Lists of topics. Adapted from 'What is this corpus about?' (Murakami et al. 2017). Reproduced with permission of the Licensor through PLSclear. 139
5.7 Three topic-lists. Adapted from 'What is this corpus about?' (Murakami et al. 2017). Reproduced with permission of the Licensor through PLSclear. 140
6.1 A possible constructicon entry for the 'responsive *for*' construction. Adapted from 'Patterns, constructions, and applied linguistics' (Hunston 2019). Reproduced with permission of John Benjamins. 150
6.2 A possible constructicon index entry. Adapted from 'Patterns, constructions, and applied linguistics' (Hunston 2019). Reproduced with permission of John Benjamins. 151
6.3 A comparative description of SkELL and #LancsBox 180
6.4 Collocates based on *research*. Adapted from 'Developing a writing assistant to help EAP writers with collocations in real time', Frankenberg-Garcia et al 2019. Reproduced with permission of the Licensor through Rightslink. 184
7.1 The dimensions identified by Gray (2015) 197

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE

Cambridge University Press & Assessment
978-1-108-42509-4 — Corpora in Applied Linguistics
Susan Hunston
Frontmatter
More Information

7.2 Self and other sources of judgement. Adapted from ‘The construction of stance in reporting clauses’ (Charles 2006). Reproduced with permission of the Licensor through Rightlink. 201
7.3 Categories of Signalling Nouns in natural sciences and social sciences. Adapted from ‘Signalling Nouns in English’ (Flowerdew and Forest 2015). Reproduced with permission of the Licensor through PLSclear. 203
7.4 Processes associated with *algorithm* in three disciplines. Adapted from ‘Scientific registers in contact’ (Teich and Holtz 2009). Reproduced with permission of John Benjamins. 206
7.5 Citations in two disciplines and one interdiscipline. Adapted from ‘Citations in interdisciplinary research articles’ (Muguiro 2020). Reproduced with permission of the Licensor through PLSclear. 207
7.6 The word family SCIENCE in four journals. Adapted from ‘Interdisciplinary Research Discourse’ (Thompson and Hunston 2020). Reproduced with permission of the Licensor through PLSclear. 208
8.1 Translations of *åka* in four languages. Adapted from ‘Seeing the lexical profile of Swedish through multilingual corpora’ (Viberg 2013). Reproduced with permission of John Benjamins. 229
8.2 Comparison between the Gettysburg Address and texts by Lincoln and Hay. Adapted from ‘Mapping lexical innovation on American social media’ (Grieve et al. 2019). 237
9.1 The ‘good at something’ construction 263
9.2 The ‘reaction at something’ construction 263
9.3 Constructions from the pattern **ADJ *for* n**. Adapted from ‘Patterns, constructions, and applied linguistics’ (Hunston 2019). Reproduced with permission of John Benjamins. 264
9.4 Constructions from the pattern **V n *of* n**. Adapted from ‘Patterns, constructions, and applied linguistics’ (Hunston 2019). Reproduced with permission of John Benjamins. 265
9.5 Levels of construction in the pattern **V *at* n**. Adapted from ‘Patterns, constructions, and applied linguistics’, Hunston 2019. Reproduced with permission of John Benjamins. 266

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE

Cambridge University Press & Assessment
978-1-108-42509-4 — Corpora in Applied Linguistics
Susan Hunston
Frontmatter
More Information

9.6 The patterns used with BARTER, EXCHANGE, SUBSTITUTE, SWAP, TRADE 282
10.1 Additions to the Coronavirus Corpus as a proportion of the NoW corpus in 2020. Adapted from ‘The Coronavirus Corpus: design, construction, and use’ (Davies 2021). 290
10.2 The comparable corpus compiled by Curry and Pérez-Paredes (2021) 291
10.3 Frequency (pmw) of FLATTEN *the curve* in the Coronavirus Corpus 296
10.4 L1 collocates of *to flatten the curve* (200 lines) 298
10.5 Verbs collocating with *vaccine confidence* and *vaccine hesitancy*, Coronavirus Corpus, April 2021 301

© in this web service Cambridge University Press & Assessment www.cambridge.org

Acknowledgements

This book owes a great deal to the influence of colleagues and fellow researchers in corpus linguistics and applied linguistics. These include friends in the Centre for Corpus Research at the University of Birmingham as well as colleagues around the world. Their support, insight and challenge have been invaluable.

In this book I draw on many of the corpus resources, both corpora and software, provided by the SketchEngine team (sketchengine.eu) and by English Corpora run by Professor Mark Davies (english-corpora.org). Other corpora and resources are acknowledged as appropriate through the book. I am grateful for permission to adapt and reproduce data of various kinds from other published sources, as acknowledged in the relevant places. I would like to mention specifically the John Benjamins series 'Studies in Corpus Linguistics' (benjamins.com/catalog/scl), and the Cambridge University Press series 'Elements in Corpus Linguistics' (www.cambridge.org/core/what-we-publish/elements/corpus-linguistics).

In the 20 years since the publication of the first edition of *Corpora in Applied Linguistics* I have had the privilege of working with PhD students, too many to mention by name, whose work has been inspirational and whose enthusiasm has always brightened my day. I would also like to acknowledge the love and continued support of friends and family, especially Ann, Jyl, Jeannette, Paul, Suganthi, Michaela, Linda and John.

© in this web service Cambridge University Press & Assessment www.cambridge.org