

The Fundamental Principles of Corpus Linguistics

How might evidence of language use – such as writing and speech – be used as a way of studying language? Corpus linguistics is the study of linguistic data from a particular language or set of languages. It is a fast-moving approach to studying language, and there is still a degree of divergence in how research questions are approached using corpus data. This book uses a framework, based on the work of Karl Popper, to explore a number of fundamental issues in corpus linguistics. It critically evaluates how these issues are tackled, and proposes a set of best practices for future research. It spells out why using corpus data is valuable, what we can learn from using it, and how we may most effectively progress our understanding of language by using such data. It is essential reading for researchers and students of language in general and of applied linguistics and English language in particular.

Tony McEnery is Distinguished Professor of Linguistics and English Language, Lancaster University and Changjiang Chair, Xi'an Jiaotong University. He has worked since the late 1980s on studying language using corpus data. He has published widely on a range of languages, topics and methods, with notable publications including *Corpus Linguistics: Method, Theory and Practice* (Cambridge University Press, 2011, with Hardie).

Vaclav Brezina is a Senior Lecturer at the Department of Linguistics and English Language and a member of the ESRC Centre for Corpus Approaches to Social Science, Lancaster University. He is interested in corpus design and methodology, and statistics. He is the author of *Statistics in Corpus Linguistics* (Cambridge University Press, 2018).

The Fundamental Principles of Corpus Linguistics

Tony McEnery

Lancaster University

Vaclav Brezina

Lancaster University



CAMBRIDGE
UNIVERSITY PRESS

Cambridge University Press & Assessment
978-1-107-04669-6 — Fundamental Principles of Corpus Linguistics
Tony McEnery, Vaclav Brezina
Frontmatter
[More Information](#)

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/9781107046696

DOI: 10.1017/9781107110625

© Tony McEnery and Vaclav Brezina 2023

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 2023

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

Library of Congress Cataloging-in-Publication Data

Names: McEnery, Tony, 1964– author. | Brezina, Vaclav, 1979– author.

Title: Fundamental principles of corpus linguistics / Tony McEnery, Lancaster University ; Vaclav Brezina.

Description: Cambridge ; New York, NY : Cambridge University Press, 2023. | Includes index.

Identifiers: LCCN 2022023390 (print) | LCCN 2022023391 (ebook) | ISBN 9781107046696 (hardback) | ISBN 9781107624689 (paperback) | ISBN 9781107110625 (ebook)

Subjects: LCSH: Corpora (Linguistics) | BISAC: LANGUAGE ARTS & DISCIPLINES / Linguistics / General

Classification: LCC P128.C68 M39 2023 (print) | LCC P128.C68 (ebook) | DDC 410.1/88–dc23/eng/20220518

LC record available at <https://lcn.loc.gov/2022023390>

LC ebook record available at <https://lcn.loc.gov/2022023391>

ISBN 978-1-107-04669-6 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.

Contents

<i>List of Figures</i>	<i>page viii</i>
<i>List of Tables</i>	<i>ix</i>
<i>Preface</i>	<i>xi</i>
Introduction	1
1 The First Sketch	3
1.1 Linguistics and Science	6
1.2 Realism and Common Sense	9
1.3 The Rational Approach to Language	13
1.4 Presuppositions	15
1.5 Logic	16
1.6 Empiricism	22
1.7 The Notion of Replication	25
1.8 Conclusion	27
2 What Is Science?	29
2.1 A Simple Model of Science	34
2.2 Problems of Induction and Verification	37
2.3 Falsifiability	42
2.4 Testing Theories	45
2.5 Science versus Rhetoric	50
2.6 What Is an Acceptable Theory?	54
2.7 Conviction, Evidence and Testing	59
2.8 The Unattainable – Truth	62
2.9 The Attainable – Provisional Truths	68
2.10 Popper and Theory	72
2.11 The Corpus and Theory	74
3 How to Do Science?	79
3.1 Consistency	82
3.2 Popper and Intuition	87
3.3 Knowing When to Stop	88
3.4 Interaction and Redundancy	91
3.5 Testability	93
3.6 Range	96

vi	Contents
3.7	The Hair Shirt of Honesty 98
3.8	Achieving Balance and Simplicity 99
3.9	Probability and Corroboration 101
3.10	Fault Lines 105
4	What Is Social Science and the Digital Humanities? 111
4.1	The Rejection of Naturalism 111
4.2	The Natural, the Social and the Cultural 113
4.3	A Focus on the Individual 119
4.4	Institutions and Traditions 120
4.5	Modelling Reality: The Oedipus Effect and the Rationality Principle 123
4.6	Critical Rationalism, Critical Realism and Social Reality 131
4.7	Propensity in Social Science 138
4.8	Conclusion 152
5	Everyday Linguistics: Form and Function 154
5.1	Prague Schools, Halliday and Axiomatic Functionalism 156
5.2	Imperfect Performance 162
5.3	The Language Learner 173
5.4	Familiarity with Context 178
5.5	Quantifying and Generalising: The Case of Wordlists 180
5.6	Conclusion 182
6	Repetition and Replication: Laying the Groundwork for an Empirical Study 184
6.1	Approaching a Study: Repetition 185
6.1.1	Ssorg's Report 190
6.1.2	The Difficulty of Repetition 194
6.2	Approaching a Study: Replication 198
6.2.1	Design of a Study 201
6.2.2	Formalising Replication 206
6.3	Final Points 212
7	Replication: Carrying out an Empirical Study 217
7.1	Method: Genuine Test and Surprise 222
7.2	Results: Hypothesis 1 and Hypothesis 2 230
7.3	Results: Hypothesis 3 and Hypothesis 4 231
7.4	Results: Hypothesis 5, Hypothesis 6 and Hypothesis 7 236
7.5	Results: Hypothesis 8 239

Contents	vii
7.6 Discussion	240
7.7 Afterglow	244
8 Conclusion	252
<i>Appendix 1: Fundamental Principles of Corpus Linguistics</i>	256
<i>Appendix 2: Glossary</i>	263
<i>References</i>	288
<i>Index</i>	307

Figures

1.1	Seismic base isolation system under the State Capitol Building in Utah	<i>page</i> 4
2.1	Popper’s tetradic scheme of problem solving	35
2.2	Black swan in Christchurch, New Zealand – home of Popper’s first academic post	38
4.1	Reality and a model	124
4.2	Forces, propensity and language	143
5.1	Utterances in learner corpus data and a normativity view	176
6.1	The population of sentences (U) in British English in the early 1990s, sampling frames/structures (A and B) and samples (two corpora, C and D)	208
6.2	The population of texts and spoken interactions (U) in British English in the early 1990s, sampling frames/structures (A and B) and two corpora, FLOB and the BNC1994 (C and D)	211

Tables

6.1	Repeating observations in the BNC1994	<i>page</i> 187
6.2	Repeatability of Leech (2003) with versions of the corpora held in CQPweb	193
6.3	The frequencies of modal verbs in four Brown family corpora, figures from Leech et al. (2010: 283) with the corresponding figures from Leech (2003) in parentheses	194
6.4	The rank of frequencies of modal verbs in four Brown family corpora	196
7.1	Comparative overview of three studies on modals in British and American English	219
7.2	Original corpora compared in Baker (2017)	222
7.3	Structure of BE06 and AmE06	223
7.4	Replication study corpora	224
7.5	Results of the replication of modals in BR. The rank of each modal in each corpus is given in parentheses after absolute frequency. The modals are ordered by the rank of the modals in FLOB.	229
7.6	Results of the replication of modals in AM. The rank of each modal in each corpus is given in parentheses after absolute frequency. The modals are ordered by the rank of the modals in Frown.	230
7.7	Frequencies and proportions of H and L modals in British and American corpora	232
7.8	Percentage loss for individual modals 1961–2006	233
7.9	Classification of modals into H, L and non-fit based on AC6	234
7.10	Frequency ranking of modals in British and American corpora	236
7.11	Spearman's correlations for ranks of modals in different corpora	238
7.12	Evaluation of the decline in frequency of individual modals between 1991/2 and 2006	239

Preface

This book started as a dialogue between the two authors, debating perspectives on corpus linguistics and its role in the exploration and expansion of the knowledge about language and society. ‘But how exactly does this work?’ we would often ask each other, coming up with quotes from philosophical literature, tentative suggestions and many more questions. We soon realised that a plausible answer to the question of the foundations of knowledge in corpus linguistics needs to transcend occasional meetings over coffee and passionate debates; indeed, a systematic answer would be required and this answer would need a book-length treatment. So, the process of writing and rewriting started, which led to a gradual crystallisation of our position and the formulation of key principles on which the argument rests. In this endeavour, we were supported by a number of people who read and commented on different parts of the draft, or who provided their subject expertise to help us with specific points. We would like to thank the following for their encouragement, generous help and useful suggestions: Charlotte Taylor, Isobel Hook, Jesse Egbert, Merja Kytö, Michael Hoey, Niall Curry, Olli Silvennoinen, Sam Kirkham, Tanja Säily and the anonymous reviewers of our work. Tony McEnery would like to express his thanks to Tony Banks who provided him with the inspiration to persevere at times. We would also like to express our gratitude to the CUP reader as well as the CUP editorial team.

This volume provides theoretical framing for systematic investigations of language with the special attention being paid to corpus linguistic methods and their epistemological underpinning. It thus builds upon the groundwork provided by two other volumes published by CUP: McEnery and Hardie (2011) *Corpus Linguistics: Method, Theory and Practice* and Brezina (2018) *Statistics in Corpus Linguistics: A Practical Guide*. The former provides the details about corpus approaches and their various traditions and applications, the latter focuses on statistical techniques used in the exploration of corpora.

This book is accompanied by a website (<http://corpora.lancs.ac.uk/stats/fundamentals.php>), which provides additional materials as well as

a printable version of the key principles to which we refer throughout this volume.

The research presented in this book was supported by the Economic and Social Research Council (grant numbers EP/P001559/1, ES/K002155/1 and ES/R008906/1).