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

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

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://lccn.loc.gov/2022023390
LC ebook record available at https://lccn.loc.gov/2022023391

ISBN 978-1-107-04669-6 Hardback

Contents

Figures

Tables

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).