Corpus Linguistics

Corpus linguistics is the study of language data on a large scale – the computer-aided analysis of very extensive collections of transcribed utterances or written texts. This textbook outlines the basic methods of corpus linguistics, explains how the discipline of corpus linguistics developed, and surveys the major approaches to the use of corpus data. It uses a broad range of examples to show how corpus data has led to methodological and theoretical innovation in linguistics in general. Clear and detailed explanations lay out the key issues of method and theory in contemporary corpus linguistics. A structured and coherent narrative links the historical development of the field to current topics in 'mainstream' linguistics. Practical activities and questions for discussion at the end of each chapter encourage students to test their understanding of what they have read and an extensive glossary provides easy access to definitions of all technical terms used in the text.

TONY MCENERY is Professor of English Language and Linguistics at Lancaster University.

ANDREW HARDIE is Lecturer in Corpus Linguistics at Lancaster University.

CAMBRIDGE TEXTBOOKS IN LINGUISTICS

Corpus Linguistics: Method, Theory and Practice

In this series:

S. C. LEVINSON *Pragmatics*
G. BROWN and G. YULE *Discourse Analysis*
R. HUDDLESTON *Introduction to the Grammar of English*
R. LASS *Phonology*
B. COMRIE *Tense*
W. KLEIN *Second Language Acquisition*
A. J. WOODS, P. FLETCHER and A. HUGHES *Statistics in Language Studies*
D. A. CRUSE *Lexical Semantics*
A. RADFORD *Transformational Grammar*
M. GARMAN *Psycholinguistics*
G. G. CORBETT *Gender*
H. J. GIEGERICH *English Phonology*
R. CANN *Formal Semantics*
J. LAVER *Principles of Phonetics*
F. R. PALMER *Grammatical Roles and Relations*
M. A. JONES *Foundations of French Syntax*
A. RADFORD *Syntactic Theory and the Structure of English: A Minimalist Approach*
R. D. VAN VALIN, JR, and R. J. LAPOLLA *Syntax: Structure, Meaning and Function*
A. DURANTI *Linguistic Anthropology*
A. CRUTTENDEN *Intonation* Second edition
J. K. CHAMBERS and P. TRUDGILL *Dialectology* Second edition
C. LYONS *Definiteness*
R. KAGER *Optimality Theory*
J. A. HOLM *An Introduction to Pidgins and Creoles*
G. G. CORBETT *Number*
C. J. EWEN and H. VAN DER HULST *The Phonological Structure of Words*
F. R. PALMER *Mood and Modality* Second edition
B. J. BLAKE *Case* Second edition
E. GUSSMAN *Phonology: Analysis and Theory*
M. YIP *Tone*
W. CROFT *Typology and Universals* Second edition
F. COULMAS *Writing Systems: An Introduction to their Linguistic Analysis*
P. J. HOPPER and E. C. TRAUGOTT *Grammaticalization* Second edition
L. WHITE *Second Language Acquisition and Universal Grammar*
I. PLAG *Word-Formation in English*
W. CROFT and A. CRUSE *Cognitive Linguistics*
A. SIEWIERSKA *Person*
A. RADFORD *Minimalist Syntax: Exploring the Structure of English*
D. BÜRING *Binding Theory*
M. BUTT *Theories of Case*
N. HORNSTEIN, J. NUÑES and K. GROHMANN *Understanding Minimalism*
B. C. LUST *Child Language: Acquisition and Growth*
G. G. CORBETT *Agreement*
J. C. L. INGRAM *Neurolinguistics: An Introduction to Spoken Language Processing and its Disorders*
J. CLACKSON *Indo-European Linguistics: An Introduction*
M. ARIEL *Pragmatics and Grammar*
R. CANN, R. KEMPSON and E. GREGOROMICHELAKI *Semantics: An Introduction to Meaning in Language*
Y. MATRAS *Language Contact*
D. BIBER and S. CONRAD *Register, Genre and Style*
L. JEFFRIES and D. MCINTYRE *Stylistics*
R. HUDSON *An Introduction to Word Grammar*
M. L. MURPHY *Lexical Meaning*
J. M. MEISEL *First and Second Language Acquisition*
T. MCENERY and A. HARDIE *Corpus Linguistics: Method, Theory and Practice*

Corpus Linguistics

Method, Theory and Practice

TONY McENERY AND
ANDREW HARDIE

Lancaster University

CAMBRIDGE
UNIVERSITY PRESS

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
Information on this title: www.cambridge.org/9780521547369

First published 2012
3rd printing 2014

Printed in the United Kingdom by Clays, St Ives plc.

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

Library of Congress Cataloguing in Publication data
McEnery, Tony, 1964–
Corpus linguistics : method, theory and practice / Tony McEnery, Andrew Hardie.
p. cm. – (Cambridge textbooks in linguistics)
Includes index.
ISBN 978-0-521-83851-1 (hardback)
1. Corpora (Linguistics) I. Hardie, Andrew. II. Title.
P128.C68M38 2011
410.1′88 – dc23 2011026519

ISBN 978-0-521-83851-1 Hardback
ISBN 978-0-521-54736-9 Paperback

Contents

Figures

Tables

Acknowledgements

This book could not have been written without the aid of many colleagues whose generous assistance we gratefully acknowledge. Svenja Adolphs, Karin Aijmer, Mark Davies, Costas Gabrielatos, Geoffrey Leech, Neil Millar and Richard Zhonghua Xiao all read part or all of the manuscript and offered useful criticisms and suggestions (although, as should go without saying, responsibility for the final text rests solely with us). We would also like to thank Eivind Torgerson for a helpful discussion regarding the similarity of corpus linguistics and sociolinguistics, and Ghada Mohamed for long talks on the topics of cluster analysis and the notion of an exhaustive linguistic feature tree (both of which fed into the discussion in Chapter 5). Others, too numerous to mention, among our colleagues and students pointed us towards relevant literature or gave us valuable sneak previews of books and research papers in press which we would not otherwise have been able to discuss here. We are thankful to them all. We are also grateful for the professionalism of the editorial staff at Cambridge University Press.

Preface

The title of this book is *Corpus Linguistics: Method, Theory and Practice*. As that title may suggest, it is about how corpus linguistics has developed and is employed as a methodology; the major theoretical issues that corpus linguists contend with today; and the problems that researchers using corpora must grapple with in practice, both within linguistics and across disciplines.

This captures in a nutshell, we like to think, what makes this book different from the many excellent introductory textbooks on corpus linguistics that have appeared since the mid-1990s. Our purpose in writing this book is not to introduce the very basics of procedures in corpus linguistics – we do not outline any step-by-step instructions describing how to go about investigating a corpus, and though we do occasionally give some example corpus analyses, we do not go into the details of how to deal with concordances, collocations, keywords and other common outputs from corpus tools. Other books have covered this ground comprehensively (e.g. Biber *et al*. 1998; Hunston 2002; Adolphs 2006; McEnery *et al*. 2006; Hoffmann *et al*. 2008), and we do not see any need to duplicate these accounts.

Instead, our aim in this textbook is to introduce, explain and in some cases problematise the most fundamental conceptual issues underlying the use of corpora, as well as reviewing what we see as the major trends of research using corpora to date. In the earlier part of the book, we will discuss corpus linguistics as a discipline, exploring high-level issues of practice that are of concern to corpus linguists, such as: features of corpus construction such as the notions of balance and representativeness; the development and exploitation of corpus tools; issues of copyright law and of research ethics; the role and limits of corpus annotation; the role of quantitative analysis; and so on. We will also review the research priorities and main contributions of a number of different schools and centres of corpus linguistics. In doing so, we will explore both their methodological apparatus and, where appropriate, the contributions to theory that they have sought to make. However, in later chapters we will move beyond the boundaries of corpus linguistics per se to consider the role that corpus methods now have across a range of types of linguistic investigation – including studies of language variation, language change, functional-cognitive linguistic theory, and psycholinguistics – and the various issues raised by the use of corpora in such enterprises. Accordingly we will argue that, although 'corpus linguistics' has clearly long had a separate existence, there is a very great degree of convergence between

corpus linguistics and these other aspects of linguistics. Corpus techniques tend no longer to be the preserve of a clearly delimited field of specialists, but rather have become a critical resource across linguistics as a whole (and beyond). Thus, we might argue that the future of the field is in 'corpus methods in linguistics' rather than 'corpus linguistics' standing independently.

This view, it must be noted, is not universally held. In particular, some scholars of the neo-Firthian school of corpus linguistics disagree entirely with it, as we will explain in our discussion of that tradition in Chapter 6. When covering matters such as this, which do not attract full consensus, we had two choices as authors. The first was to attempt to maintain neutrality, and to write the text without letting our opinions and theoretical and methodological preferences colour the account. The second was to acknowledge our perspective, and associated biases, frankly and explicitly, and to provide an account in which we explain and justify our views as best we can. It is this second approach which we have adopted. Our discussion of the different traditions of corpus research, of the wider use of corpora in linguistics and beyond, and of the future directions we see as desirable, thus amount to what might be called a position statement for our 'version' of corpus linguistics (to borrow the memorable phrasing of Teubert 2005). We make no pretence to neutrality, and our discussions of traditions in corpus linguistics other than our own may indeed characterise a given school of thought in terms that the researchers within that tradition would not necessarily agree with. The most notable case of this is that our discussion of neo-Firthian corpus linguistics is very much informed by our stance as non-neo-Firthian researchers. We urge the reader who is interested in understanding the full picture with regard to these debates to refer to accounts in which the scholars we discuss have dealt with these issues from their own perspective, with their different sets of preconceptions, opinions and biases; where appropriate we list such accounts in the suggestions for further reading provided in each chapter. In the case of neo-Firthian theory, for instance, we would not hesitate to recommend Tognini-Bonelli (2001) or Teubert and Čermáková (2004) as clear and readable presentations of the 'other side' of the argument that we make here in Chapters 6 and 8.

Our recommendations for further reading cover both the primary and the secondary literature, as appropriate. We have also included some study questions at the end of each chapter. These are divided into two groups. For readers who are interested in thinking further about some of the issues and problems that we cover, we suggest a number of *questions for discussion*. Alongside that, we also provide some *practical activities*. These activities provide practice in some of the key procedures that are needed in actually 'doing' corpus linguistics. All the practical activities assume that you have access to a reasonably large corpus and some corpus analysis software, but do not require any specific corpus. So if you have a copy of a suitable corpus on your computer, you can work with that data using any one of several different corpus tools to complete the exercise (some of which are available for free on the Internet). If not, then a number

of large, standard corpora – in several languages – can be accessed and analysed via the World Wide Web, using online interfaces which either are openly accessible, or offer freely available sign-in accounts. There are too many of these to list here (but see our discussion in Chapter 2). For English, we recommend BNCweb (http://bncweb.lancs.ac.uk/bncwebSignup) or the Brigham Young University online interface (http://corpus.byu.edu). While different software tools have different capabilities, the activities in this book are based on the core functions that all corpus tools make available.

The centrality of the Internet to the practice of corpus linguistics today means that it was necessary to make reference to websites and web services at various points in the book. We have tried to keep the number of web addresses in the text to a minimum, in the light of how variable these can be over relatively short periods of time. It is inevitable, however, that some of these web addresses will become outdated over time. For this reason, we have established a companion website for this book, where updated links will be made available where necessary. The website will also contain other supplementary material, including suggested answers to the study questions in each chapter. Importantly, this site also contains a large number of additional notes that were simply too numerous to include in the book. Accordingly, we suggest that you check the website after you read each chapter. The companion website address is www.cambridge.org/mcenery-hardie.