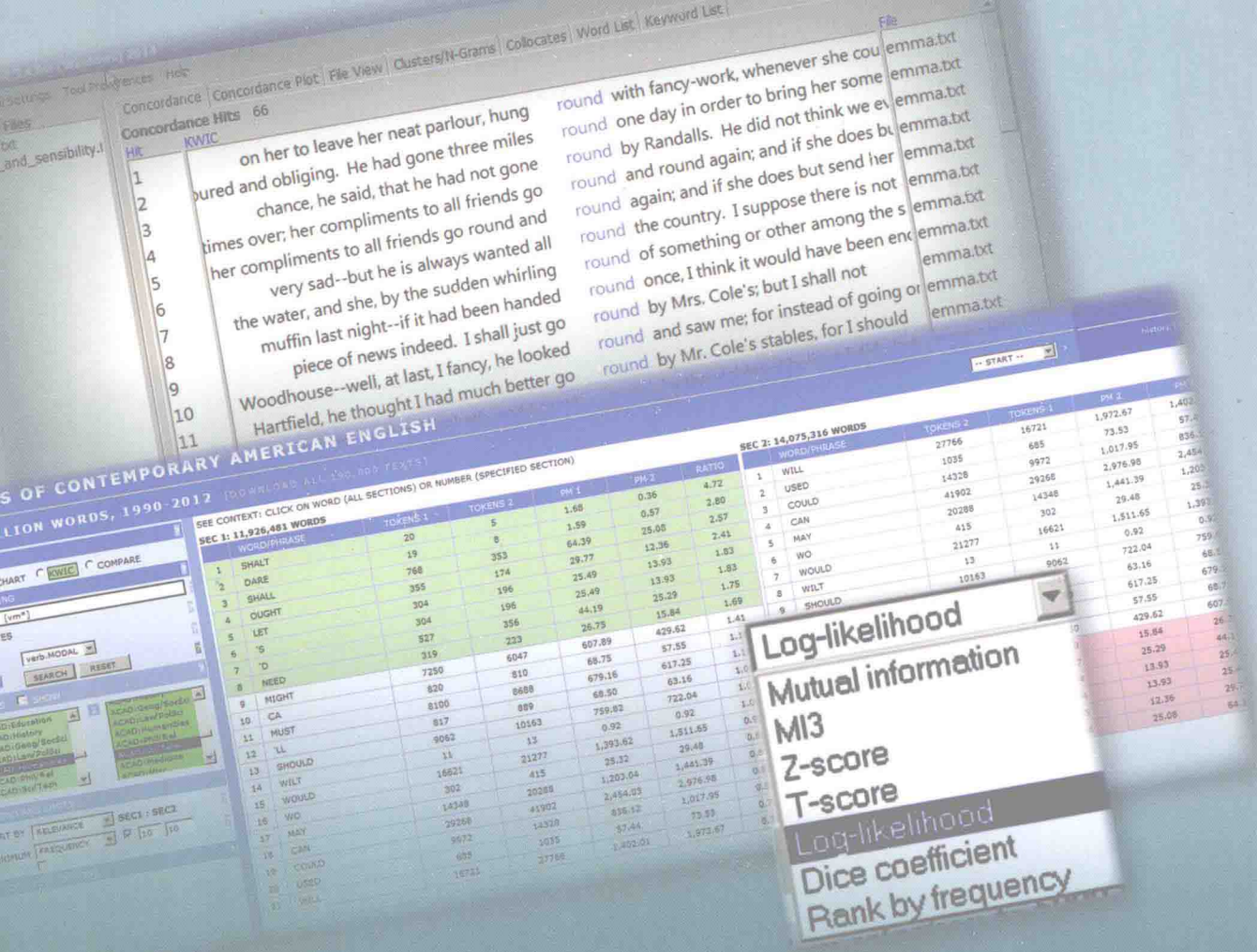

PRACTICAL

Corpus Linguistics

An Introduction to Corpus-Based Language Analysis

MARTIN WEISSER

with website

WILEY Blackwell

Practical Corpus Linguistics

An Introduction to Corpus-Based Language Analysis

Martin Weisser

WILEY Blackwell

This edition first published 2016

Registered Office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Offices
350 Main Street, Malden, MA 02148-5020, USA
9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, for customer services, and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

Library of Congress Cataloging-in-Publication Data

Weisser, Martin, author.
Practical corpus linguistics : an introduction to corpus-based language analysis / Martin Weisser. – First edition.
pages cm
Includes bibliographical references and index.
ISBN 978-1-118-83187-8 (hardback) – ISBN 978-1-118-83188-5 (paper) 1. Linguistic analysis (Linguistics)–Databases. 2. Linguistic analysis (Linguistics)–Software. 3. Corpora (Linguistics)–Methodology. 4. Corpora (Linguistics)–Technological innovations. 5. Computational linguisitics–Methodology. 6. Computer network resources–Evaluation. 7. Citation of electronic information resources. I. Title.
P128.D37W45 2016
410.1′88–dc23

2015023709

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

Cover image: [Production Editor to insert]

Set in 10.5/13pt Galliard by Aptara Inc., New Delhi, India

Printed in Singapore by C.O.S. Printers Pte Ltd

1 2016

Practical Corpus Linguistics

To Ye & Emma,
who've had to suffer
from an undue lack of attention
throughout the final months
of writing this book

List of Figures

List of Tables

Acknowledgements

I'd first like to start by thanking my former students in Chemnitz, Bayreuth and Hong Kong who 'suffered through' the initial sets of teaching materials that eventually formed the basis for writing this textbook. The next big thanks needs to go to my colleagues here at Guangdong University of Foreign Studies, who attended a series of workshops where I tested out the materials from the preliminary drafts of several chapters and who provided me with highly useful feedback. Particular mention here deserves to go to Junyu (Mike) ZHANG, who not only commented on the content of several chapters, but also pointed out certain issues of style that have hopefully helped me to make the writing more accessible to an international readership.

My next round of thank yous goes to Yanping DONG and Hai XU, for allowing me to join the National Key Research Center for Linguistics and Applied Linguistics at Guangdong University of Foreign Studies, which has provided me with more of the desperately needed time to focus on writing this book, while also allowing me to conduct other types of research that have influenced the contents of the book in various ways. To Hai XU, I give additional thanks for sharing his experience in, and knowledge of, corpora of Chinese, which has, unfortunately, only partially found its way into this book, due to limits of space. To my other colleagues, especially Yiqiong ZHANG, I also give thanks for providing a more moral type of support through engaging in further discussions and making me feel at home in the Center.

I also owe a great debt to Laurence Anthony for allowing me to pester him with a series of questions about and suggestions for improving AntConc. More or less the same, though possibly to a slightly lesser extent, goes for Sebastian Hoffmann and Mark Davies for answering questions about particular features of, and again

partly responding to requests for improving, BNCweb and the COCA interface, respectively.

Next, I'd sincerely like to thank the anonymous reviewers of this book, who, through their many invaluable constructive comments, have not only encouraged me in writing the book, but also hopefully allowed me to improve the contents substantially.

My final – but most important and heartfelt – note of thanks goes to Geoff Leech. Although credit for introducing me to the study of corpus linguistics has to go to someone else, he's certainly been the single most important influence on my career and thinking as a corpus linguist. I'll forever be grateful for his ongoing support throughout my years spent at Lancaster, working with him, as external examiner of my PhD, and later up to his untimely demise in August 2014. I sincerely hope that he would have appreciated the design and critical aims of this textbook, and perhaps also recognised his implicit hand in shaping it…

Contents

1

Introduction

This textbook aims to teach you how to analyse and interpret language data in written or orthographically transcribed form (i.e. represented as if it were written, if the original data is spoken). It will do so in a way that should not only provide you with the technical skills for such an analysis for your own research purposes, but also raise your awareness of how corpus evidence can be used in order to develop a better understanding of the forms and functions of language. It will also teach you how to use corpus data in more applied contexts, such as e.g. in identifying suitable materials/examples for language teaching, investigating socio-linguistic phenomena, or even trying to verify existing linguistic theories, as well as to develop your own hypotheses about the many different aspects of language that can be investigated through corpora. The focus will primarily be on English-language data, although we may occasionally, whenever appropriate, refer to issues that could be relevant to the analysis of other languages. In doing so, we'll try to stay as theory-neutral as possible, so that no matter which 'flavour(s)' of linguistics you may have been exposed to before, you should always be able to understand the background to all the exercises or questions presented here.

The book is aimed at a variety of readers, ranging mainly from linguistics students at senior undergraduate, Masters, or even PhD levels who are still unfamiliar with corpus linguistics, to language teachers or textbook developers who want to create or employ more real-life teaching materials. As many of the techniques we'll be dealing with here also allow us to investigate issues of style in both literary and non-literary text, and much of the data we'll initially use actually consists of fictional works because these are easier to obtain and often don't cause any copyright

Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis, First Edition. Martin Weisser.

issues, the book should hopefully also be useful to students of literary stylistics. To some extent, I also hope it may be beneficial to computer scientists working on language processing tasks, who, at least in my experience, often lack some crucial knowledge in understanding the complexities and intricacies of language, and frequently tend to resort to mathematical methods when more linguistic (symbolic) ones would be more appropriate, even if these may make the process of writing 'elegant' and efficient algorithms more difficult.

You may also be asking yourself why you should still be using a textbook at all in this day and age, when there are so many video tutorials available, and most programs offer at least some sort of online help to get you started. Essentially, there are two main reasons for this: a) such sources of information are only designed to provide you with a basic overview, but don't actually teach you, simply demonstrating how things are done. In other words they may do a relatively good job in showing you one or more ways of doing a few things, but often don't really allow you to use a particular program independently and for more complex tasks than the author of the tutorial/help file may actually have envisaged. And b) online tutorials, such as the ones on YouTube, may not only take a rather long time to (down)load, but might not even be (easily) accessible in some parts of the world at all, due to internet censorship.

If you're completely new to data analysis on the computer and working with – as opposed to simply opening and reading – different file types, some of the concepts and methods we'll discuss here may occasionally make you feel like you're doing computer science instead of working with language. This is, unfortunately, something you'll need to try and get used to, until you begin to understand the intricacies of working with language data on the computer better, and, by doing so, will also develop your understanding of the complexity inherent in language (data) itself. This is by no means an easy task, so working with this book, and thereby trying to develop a more complete understanding of language and how we can best analyse and describe it, be it for linguistic or language teaching purposes, will often require us to do some very careful reading and thinking about the points under discussion, so as to be able to develop and verify our own hypotheses about particular language features. However, doing so is well worth it, as you'll hopefully realise long before reaching the end of the book, as it opens up possibilities for understanding language that go far beyond a simple manual, small-scale, analysis of texts.

In order to achieve the aims of the book, we'll begin by discussing which types of data are already readily available, exploring ways of obtaining our own data, and developing an understanding of the nature of electronic documents and what may make them different from the more traditional types of printed documents we're all familiar with. This understanding will be developed further throughout the book, as we take a look at a number of computer programs that will help us to conduct our analyses at various levels, ranging from words to phrases, and to even larger units of text. At the same time, of course, we cannot ignore the fact that there may be issues in corpus linguistics related to lower levels, such