



PRACTICAL

Corpus Linguistics

An Introduction to Corpus-Based Language Analysis

MARTIN WEISSER



WILEY Blackwell

Practical Corpus Linguistics

An Introduction to Corpus Based
Language Analysis

Martin Weisser

WILEY Blackwell

This edition first published 2016
© 2016 John Wiley & Sons, Inc

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.

The right of Martin Weisser to be identified as the author of this work has been asserted in accordance with the UK Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and authors have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

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 linguistics–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

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

3.1	Illustration of basic document structure	34
4.1	The ICEweb interface	50
5.1	Example of a KWIC concordance output	69
5.2	The AntConc startup screen	70
5.3	AntConc file opening options	71
5.4	AntConc file settings	71
5.5	AntConc ‘Corpus Files’ window (two files loaded)	72
5.6	AntConc ‘Search Term’ and search options	72
5.7	AntConc results for <i>round</i> in two novels by Jane Austen	73
5.8	AntConc ‘Search Window Size’ options	73
5.9	AntConc ‘Kwic Sort’ options	74
6.1	Sample paragraph for practising and understanding regex patterns	84
7.1	Sample output of the Simple PoS Tagger	110
8.1	The BNCweb startup screen	122
8.2	Results for simple search for <i>assume</i>	123
8.3	BNCweb query follow-on options	125
8.4	The basic COCA interface	133
8.5	Display of antonyms <i>thoughtful</i> and <i>thoughtless</i> as alternatives	134
8.6	Side-by-side comparison for the lemma of <i>movie</i> in the COCA and BNC	136
9.1	Output of a basic frequency list in AntConc	152
9.2	Token (word) (re-)definition in AntConc	154
9.3	AntConc Word List preferences	159
9.4	BNCweb frequency list selection options	161
9.5	Options for defining subcorpora in BNCweb	162
9.6	Excel text import wizard (stage 1)	164

9.7	Excel text import wizard (stage 3)	165
9.8	Sort options in Excel	166
9.9	Options for defining subcorpora according to genre	167
9.10	BNCweb keyword and title scan	168
9.11	AntConc Keyword preferences	170
9.12	Keyword options in BNCweb	172
9.13	Keyword comparison of university essays and written component of the BNC (top 31 entries)	173
10.1	Illustration of a collocation span	208
10.2	Options for statistical collocation measures in BNCweb	209
11.1	A brief SGML sample	232
11.2	CSS sample paragraph styling	235
11.3	TEI header for BNC file KST	247

List of Tables

2.1	Extract from <i>Beowulf</i> , encoded/represented in two different ways	14
2.2	Early written corpora	16
2.3	Composition of the Brown Corpus	17
2.4	Some examples of (earlier) spoken corpora	18
2.5	Early mixed corpora	19
2.6	Modern mega corpora	20
2.7	Examples of diachronic corpora	21
2.8	Examples of academic corpora	22
2.9	Examples of learner corpora	23
2.10	Selection of pragmatically annotated corpora	24
3.1	Common file formats and their properties	37
7.1	Ambiguous PoS tags in the Brown Corpus	102
7.2	The Penn Treebank tagset (based on Taylor et al. 2003: 8)	103
7.3	The CLAWS 7 (C7) tagset	105
8.1	Wildcards and their uses for investigating linguistic features	129
8.2	Simplified tags in BNCweb	132
8.3	Word + PoS tags breakdown for <i>mind</i>	142
9.1	Top 15 most frequent word types in section A of the LOB Corpus	157
9.2	Recalculated norming sample from Biber, Conrad & Reppen (1998: 263)	175
11.1	Annotation types listed in Garside et al. 1997	228
11.2	CSS properties for XML visualisation exercise	240
11.3	Colour semantics and CSS styles	243

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

List of Figures	xiii
List of Tables	xv
Acknowledgements	xvii
1 Introduction	1
1.1 Linguistic Data Analysis	3
1.1.1 What's data?	3
1.1.2 Forms of data	3
1.1.3 Collecting and analysing data	7
1.2 Outline of the Book	8
1.3 Conventions Used in this Book	10
1.4 A Note for Teachers	11
1.5 Online Resources	11
2 What's Out There?	13
2.1 What's a Corpus?	13
2.2 Corpus Formats	13
2.3 Synchronic vs. Diachronic Corpora	15
2.3.1 'Early' synchronic corpora	15
2.3.2 Mixed corpora	18
2.3.3 Examples of diachronic corpora	20
2.4 General vs. Specific Corpora	21
2.4.1 Examples of specific corpora	22
2.5 Static Versus Dynamic Corpora	25
2.6 Other Sources for Corpora	26

Solutions to/Comments on the Exercises	26
Note	28
Sources and Further Reading	28
3 Understanding Corpus Design	29
3.1 Food for Thought – General Issues in Corpus Design	29
3.1.1 Sampling	30
3.1.2 Size	31
3.1.3 Balance and representativeness	32
3.1.4 Legal issues	32
3.2 What's in a Text? – Understanding Document Structure	33
3.2.1 Headers, 'footers' and meta-data	34
3.2.2 The structure of the (text) body	36
3.2.3 What's (in) an electronic text? – understanding file formats and their properties	37
3.3 Understanding Encoding: Character Sets, File Size, etc.	38
3.3.1 ASCII and legacy encodings	38
3.3.2 Unicode	39
3.3.3 File sizes	40
Solutions to/Comments on the Exercises	41
Sources and Further Reading	42
4 Finding and Preparing Your Data	43
4.1 Finding Suitable Materials for Analysis	44
4.1.1 Retrieving data from text archives	44
4.1.2 Obtaining materials from Project Gutenberg	44
4.1.3 Obtaining materials from the Oxford Text Archive	45
4.2 Collecting Written Materials Yourself ('Web as Corpus')	46
4.2.1 A brief note on plain-text editors	46
4.2.2 Browser text export	48
4.2.3 Browser HTML export	49
4.2.4 Getting web data using ICEweb	50
4.2.5 Downloading other types of files	52
4.3 Collecting Spoken Data	53
4.4 Preparing Written Data for Analysis	56
4.4.1 'Cleaning up' your data	56
4.4.2 Extracting text from proprietary document formats	58
4.4.3 Removing unnecessary header and 'footer' information	58
4.4.4 Documenting what you've collected	59
4.4.5 Preparing your data for distribution or archiving	60
Solutions to/Comments on the Exercises	62
Sources and Further Reading	66
5 Concordancing	67
5.1 What's Concordancing?	67

5.2	Concordancing with AntConc	69
5.2.1	Sorting results	74
5.2.2	Saving, pruning and reusing your results	75
	Solutions to/Comments on the Exercises	78
	Sources and Further Reading	81
6	Regular Expressions	82
6.1	Character Classes	84
6.2	Negative Character Classes	86
6.3	Quantification	86
6.4	Anchoring, Grouping and Alternation	87
6.4.1	Anchoring	87
6.4.2	Grouping and alternation	88
6.4.3	Quoting and using special characters	90
6.4.4	Constraining the context further	91
6.5	Further Exercises	92
	Solutions to/Comments on the Exercises	93
	Sources and Further Reading	100
7	Understanding Part-of-Speech Tagging and Its Uses	101
7.1	A Brief Introduction to (Morpho-Syntactic) Tagsets	103
7.2	Tagging Your Own Data	109
	Solutions to/Comments on the Exercises	113
	Sources and Further Reading	120
8	Using Online Interfaces to Query Mega Corpora	121
8.1	Searching the BNC with BNCweb	122
8.1.1	What is BNCweb?	122
8.1.2	Basic standard queries	123
8.1.3	Navigating through and exploring search results	124
8.1.4	More advanced standard query options	126
8.1.5	Wildcards	126
8.1.6	Word and phrase alternation	128
8.1.7	Restricting searches through PoS tags	129
8.1.8	Headword and lemma queries	131
8.2	Exploring COCA through the BYU Web-Interface	132
8.2.1	The basic syntax	133
8.2.2	Comparing corpora in the BYU interface	135
	Solutions to/Comments on the Exercises	137
	Sources and Further Reading	145
9	Basic Frequency Analysis – or What Can (Single) Words Tell Us About Texts?	146
9.1	Understanding Basic Units in Texts	146
9.1.1	What's a word?	147
9.1.2	Types and tokens	149

9.2	Word (Frequency) Lists in AntConc	151
9.2.1	Stop words – good or bad?	156
9.2.2	Defining and using stop words in AntConc	158
9.3	Word Lists in BNCweb	160
9.3.1	Standard options	160
9.3.2	Investigating subcorpora	162
9.3.3	Keyword lists	169
9.4	Keyword Lists in AntConc and BNCweb	169
9.4.1	Keyword lists in AntConc	169
9.4.2	Keyword lists in BNCweb	172
9.5	Comparing and Reporting Frequency Counts	175
9.6	Investigating Genre-Specific Distributions in COCA	178
	Solutions to/Comments on the Exercises	179
	Sources and Further Reading	192
10	Exploring Words in Context	193
10.1	Understanding Extended Units of Text	194
10.2	Text Segmentation	195
10.3	N-Grams, Word Clusters and Lexical Bundles	196
10.4	Exploring (Relatively) Fixed Sequences in BNCweb	198
10.5	Simple, Sequential Collocations and Colligations	198
10.5.1	‘Simple’ collocations	198
10.5.2	Colligations	200
10.5.3	Contextually constrained and proximity searches	201
10.6	Exploring Colligations in COCA	202
10.7	N-grams and Clusters in AntConc	205
10.8	Investigating Collocations Based on Statistical Measures in AntConc, BNCweb and COCA	207
10.8.1	Calculating collocations	207
10.8.2	Computing collocations in AntConc	209
10.8.3	Computing collocations in BNCweb	210
10.8.4	Computing collocations in COCA	211
	Solutions to/Comments on the Exercises	212
	Sources and Further Reading	226
11	Understanding Markup and Annotation	227
11.1	From SGML to XML – A Brief Timeline	229
11.2	XML for Linguistics	230
11.2.1	Why bother?	230
11.2.2	What does markup/annotation look like?	230
11.2.3	The ‘history’ and development of (linguistic) markup	232
11.2.4	XML and style sheets	234
11.3	‘Simple XML’ for Linguistic Annotation	236
11.4	Colour Coding and Visualisation	240
11.5	More Complex Forms of Annotation	246

Solutions to/Comments on the Exercises	248
Sources and Further Reading	253
12 Conclusion and Further Perspectives	254
Appendix A: The CLAWS C5 Tagset	259
Appendix B: The Annotated Dialogue File	261
Appendix C: The CSS Style Sheet	269
Glossary	271
References	277
Index	283

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

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