

Contents

Acknowledgments

viii

1. Introduction	1
1.1 Why Another Introduction to Corpus Linguistics?	1
1.2 Outline of the Book	4
1.3 Recommendation for Instructors	5
2. The Three Central Corpus-linguistic Methods	7
2.1 Corpora	7
2.1.1 What is a Corpus?	7
2.1.2 What Kinds of Corpora are There?	9
2.2 Frequency Lists	12
2.3 Lexical Co-occurrence: Collocations	14
2.4 (Lexico-)Grammatical Co-occurrence: Concordances	16
3. An Introduction to R	19
3.1 A Few Central Notions: Data Structures, Functions, and Arguments	23
3.2 Vectors	28
3.2.1 Basics	28
3.2.2 Loading Vectors	32
3.2.3 Accessing and Processing (Parts of) Vectors	35
3.2.4 Saving Vectors	42
3.3 Factors	43
3.4 Data Frames	44
3.4.1 Generating Data Frames	44
3.4.2 Loading and Saving Data Frames	46
3.4.3 Accessing and Processing (Parts of) Data Frames	48
3.5 Lists	53
3.6 Elementary Programming Functions	59
3.6.1 Conditional Expressions	59
3.6.2 Loops	60
3.6.3 Rules of Programming	64
3.7 Character/String Processing	68
3.7.1 Getting Information from and Accessing (Vectors of) Character Strings	69
3.7.2 Elementary Ways to Change (Vectors of) Character Strings	70
3.7.3 Merging and Splitting (Vectors of) Character Strings without Regular Expressions	70
3.7.4 Searching and Replacing without Regular Expressions	72
3.7.5 Searching and Replacing with Regular Expressions	79

3.7.6	Merging and Splitting (Vectors of) Character Strings with Regular Expressions	96
3.8	File and Directory Operations	99
4.	Using R in Corpus Linguistics	105
4.1	Frequency Lists	106
4.1.1	A Frequency List of an Unannotated Corpus	106
4.1.2	A Reverse Frequency List of an Unannotated Corpus	110
4.1.3	A Frequency List of an Annotated Corpus	112
4.1.4	A Frequency List of Tag-word Sequences from an Annotated Corpus	114
4.1.5	A Frequency List of Word Pairs from an Annotated Corpus	118
4.1.6	A Frequency List of an Annotated Corpus (with One Word Per Line)	124
4.1.7	A Frequency List of Word Pairs of an Annotated Corpus (with One Word Per Line)	126
4.2	Concordances	127
4.2.1	A Concordance of an Unannotated Text File	127
4.2.2	A Simple Concordance from Files of a POS-tagged (SGML) Corpus	135
4.2.3	More Complex Concordances from Files of a POS-tagged (SGML) Corpus	141
4.2.4	A Lemma-based Concordance from Files of a POS-tagged and Lemmatized (XML) Corpus	146
4.3	Collocations	149
4.4	Excursus 1: Processing Multi-tiered Corpora	156
4.5	Excursus 2: Unicode	166
4.5.1	Frequency Lists	167
4.5.2	Concordancing	169
5.	Some Statistics for Corpus Linguistics	173
5.1	Introduction to Statistical Thinking	174
5.1.1	Variables and their Roles in an Analysis	174
5.1.2	Variables and their Information Value	174
5.1.3	Hypotheses: Formulation and Operationalization	176
5.1.4	Data Analysis	182
5.1.5	Hypothesis (and Significance) Testing	183
5.2	Categorical Dependent Variables	189
5.2.1	One Categorical Dependent Variable, No Independent Variable	189
5.2.2	One Categorical Dependent Variable, One Categorical Independent Variable	192
5.2.3	One Categorical Dependent Variable, 2+ Independent Variables	200
5.3	Interval/Ratio-scaled Dependent Variables	201
5.3.1	Descriptive Statistics for Interval/Ratio-scaled Dependent Variables	201
5.3.2	One Interval/Ratio-scaled Dependent Variable, One Categorical Independent Variable	205
5.3.3	One Interval/Ratio-scaled Dependent Variable, One Interval/Ratio-scaled Independent Variable	211
5.3.4	One Interval/Ratio-scaled Dependent Variable, 2+ Independent Variables	214
5.4	Customizing Statistical Plots	215
5.5	Reporting Results	215

6. Case Studies and Pointers to Other Applications	219
6.1 Introduction to the Case Studies	219
6.2 Some Pointers to Further Applications	220
Appendix	225
References	229
Endnotes	237
Index	243

Introduction

I wish to thank the following people for input and support: Steven J. Chung, James S. Danks, Laura Kassar, and Gillian Law for going over large parts of the book and making and much appreciated suggestions for improvement; and Caroline V. David for discussion of an early draft of this book. Also I am particularly grateful to two of the reviewers, one from the Graduate Student Reviewer and one who turned out to be Haridh Rajaram, as well as to many students and participants of classes, summer schools, and workshops who have provided me with feedback for the last few years. Also, I thank 'Laptop' for his input and much valuable input and discussion at work as well as for his and their support for input regarding regular expressions and HTML. I am also grateful to many students who have provided me with feedback and input. I think the book is radically different from every other introduction in corpus linguistics that I have seen. For example, there are a lot of things that are regularly dealt with in introductions to corpus linguistics that I will only be concerned with very briefly:

- the history of corpus linguistics: Keating, Tyler, early (no word corpora), up to the contemporary 'large corpora' and the lively web-as-corpus discussion;
- how to sample corpora: size, sampling, balancedness, representativity, ...
- how to create corpora: marking and annotation: lemmatization, tagging, parsing, ...
- types and examples of corpora: synchronic vs. diachronic, annotated vs. unannotated;
- what kinds of corpus-linguistic research other people have done.

One important characteristic of this book is that I would like to teach you how to do corpus linguistics. This is important since, as we know, corpus linguistics is a method of analysis and thus talking about how to do things should enjoy a high priority. Therefore, as opposed to reporting many previous studies, I will be more concerned with:

- aspects of this method: how to generate frequency lists, concordances, collocation displays, etc. (where I will not know these terms, they will be explained later);
- aspects of data extraction: how to import data into a spreadsheet program for further processing; how to analyze results statistically; how to represent the results properly; how to report your results.

A second important characteristic of this book is that it only uses open source software: