# 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