Contents

List of Figures xi
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
## CONTENT

Solutions to/Comments on the Exercises  
Note  
Sources and Further Reading  

### 3 Understanding Corpus Design

#### 3.1 Food for Thought – General Issues in Corpus Design
  
  3.1.1 Sampling  
  3.1.2 Size  
  3.1.3 Balance and representativeness  
  3.1.4 Legal issues  

#### 3.2 What’s in a Text? – Understanding Document Structure
  
  3.2.1 Headers, ‘footers’ and meta-data  
  3.2.2 The structure of the (text) body  
  3.2.3 What’s (in) an electronic text? – understanding file formats and their properties  

#### 3.3 Understanding Encoding: Character Sets, File Size, etc.
  
  3.3.1 ASCII and legacy encodings  
  3.3.2 Unicode  
  3.3.3 File sizes  

### 4 Finding and Preparing Your Data

#### 4.1 Finding Suitable Materials for Analysis
  
  4.1.1 Retrieving data from text archives  
  4.1.2 Obtaining materials from Project Gutenberg  
  4.1.3 Obtaining materials from the Oxford Text Archive  

#### 4.2 Collecting Written Materials Yourself (‘Web as Corpus’)  
  
  4.2.1 A brief note on plain-text editors  
  4.2.2 Browser text export  
  4.2.3 Browser HTML export  
  4.2.4 Getting web data using ICEweb  
  4.2.5 Downloading other types of files  

#### 4.3 Collecting Spoken Data  

#### 4.4 Preparing Written Data for Analysis
  
  4.4.1 ‘Cleaning up’ your data  
  4.4.2 Extracting text from proprietary document formats  
  4.4.3 Removing unnecessary header and ‘footer’ information  
  4.4.4 Documenting what you’ve collected  
  4.4.5 Preparing your data for distribution or archiving  

### 5 Concordancing

#### 5.1 What’s Concordancing?
9.2 Word (Frequency) Lists in AntConc
  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
  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
  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
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
CONTENTS

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