

# Contents

<i>List of Figures</i>	<i>page x</i>
<i>List of Tables</i>	<i>xiv</i>
<i>About This Book</i>	<i>xvii</i>
<i>Acknowledgements</i>	<i>xix</i>
<b>1 Introduction: Statistics Meets Corpus Linguistics</b>	<b>1</b>
1.1 What Is This Chapter About?	1
1.2 What Is Statistics? Science, Corpus Linguistics and Statistics	1
1.3 Basic Statistical Terminology	5
1.4 Building of Corpora and Research Design	15
1.5 Exploring Data and Data Visualization	22
1.6 Application and Further Examples: Do Fiction Writers Use More Adjectives than Academics?	30
1.7 Exercises	32
Things to Remember	36
Advanced Reading	36
<b>2 Vocabulary: Frequency, Dispersion and Diversity</b>	<b>38</b>
2.1 What Is This Chapter About?	38
2.2 Tokens, Types, Lemmas and Lexemes	38
2.3 Words in a Frequency List	42
2.4 The Whelk Problem: Dispersion	46
2.5 Which Words Are Important? Average Reduced Frequency	53
2.6 Lexical Diversity: Type/Token Ratio (TTR), STTR and MATTR	57
2.7 Application and Further Examples: Do the British Talk about Weather All the Time?	59
2.8 Exercises	62
Things to Remember	64
Advanced Reading	65
<b>3 Semantics and Discourse: Collocations, Keywords and Reliability of Manual Coding</b>	<b>66</b>
3.1 What Is This Chapter About?	66
3.2 Collocations and Association Measures	66
3.3 Collocation Graphs and Networks: Exploring Cross-associations	75
3.4 Keywords and Lockwords	79
3.5 Inter-rater Agreement Measures	87

---

3.6 Application and Further Examples: What Do Readers of British Newspapers Think about Immigration?	92
3.7 Exercises	96
Things to Remember	100
Advanced Reading	101
<b>4 Lexico-grammar: From Simple Counts to Complex Models</b>	<b>102</b>
4.1 What Is This Chapter About?	102
4.2 Analysing a Lexico-grammatical Feature	103
4.3 Cross-tabulation, Percentages and Chi-squared Test	108
4.4 Logistic Regression	117
4.5 Application: <i>That</i> or <i>Which</i> ?	130
4.6 Exercises	134
Things to Remember	137
Advanced Reading	138
<b>5 Register Variation: Correlation, Clusters and Factors</b>	<b>139</b>
5.1 What Is This Chapter About?	139
5.2 Relationships between Variables: Correlations	139
5.3 Classification: Hierarchical Agglomerative Cluster Analysis	151
5.4 Multidimensional Analysis (MD)	160
5.5 Application: Registers in New Zealand English	170
5.6 Exercises	177
Things to Remember	181
Advanced Reading	182
<b>6 Sociolinguistics and Stylistics: Individual and Social Variation</b>	<b>183</b>
6.1 What Is This Chapter About?	183
6.2 Individual Style and Social Variation: Where Does a Sociolinguistic Variable Start?	183
6.3 Group Comparison: T-Test, ANOVA, Mann–Whitney <i>U</i> Test, Kruskal–Wallis Test	186
6.4 Individual Style: Correspondence Analysis	199
6.5 Linguistic Context: Mixed-Effects Models	207
6.6 Application: Who Is This Person from the White House?	211
6.7 Exercises	215
Things to Remember	217
Advanced Reading	218
<b>7 Change over Time: Working Diachronic Data</b>	<b>219</b>
7.1 What Is This Chapter About?	219
7.2 Time as a Variable: Measuring and Visualizing Time	219

---

7.3 Finding and Interpreting Differences: Percentage Change and the Bootstrap Test	229
7.4 Grouping Time Periods: Neighbouring Cluster Analysis	235
7.5 Modelling Changes in Discourse: Peaks and Troughs and UFA	241
7.6 Application: Colours in the Seventeenth Century	247
7.7 Exercises	251
Things to Remember	255
Advanced Reading	256
<b>8 Bringing Everything Together: Ten Principles of Statistical Thinking, Meta-analysis and Effect Sizes</b>	<b>257</b>
8.1 What Is This Chapter About?	257
8.2 Ten Principles of Statistical Thinking	257
8.3 Meta-analysis: Statistical Synthesis of Research Results	267
8.4 Effect Sizes: A Guide for Meaningful Use	275
8.5 Exercises	280
Things to Remember	282
Advanced Reading	282
<i>Final Remarks</i>	283
<i>References</i>	285
<i>Index</i>	294