Contents

Preface xii
Acknowledgements xiii
Abbreviations xv

1 Basic statistics 1
- 1 Introduction 1
- 2 Describing data 2
  - 2.1 Measures of central tendency 2
  - 2.2 Probability theory and the normal distribution 3
  - 2.3 Measures of variability 6
  - 2.4 The *z* score 7
  - 2.5 Hypothesis testing 9
  - 2.6 Sampling 9
- 3 Comparing groups 10
  - 3.1 Parametric versus non-parametric procedures 10
  - 3.2 Parametric comparison of two groups 11
    - 3,2.1 The *t* test for independent samples 11
    - 3.2.2 Use of the *t* test in corpus linguistics 14
    - 3.2.3 The matched pairs *t* test 15
    - 3.2.4 Use of the matched pairs *t* test in corpus linguistics 16
  - 3.3 Non-parametric comparisons of two groups 16
    - 3.3.1 The Wilcoxon rank sums test or Mann–Whitney *U* test 17
    - 3.3.2 Use of the Wilcoxon rank sums test in corpus linguistics 17
    - 3.3.3 The median test 18
    - 3.3.4 Use of the median test in corpus linguistics 19
    - 3.3.5 Non-parametric comparisons: repeated measures 20
      - 3.3.5.1 The sign test 20
      - 3.3.5.2 Use of the sign test in corpus linguistics 20

3.3.5.3 The Wilcoxon matched pairs signed ranks test 21
3.3.5.4 A linguistic application of the Wilcoxon matched pairs signed ranks test 21
3.4 Comparisons between three or more groups 22
3.4.1 Analysis of variance (ANOVA) 22
3.4.2 Use of ANOVA in corpus linguistics 23
4 Describing relationships 24
4.1 The chi-square test 24
4.2 Use of the chi-square test in corpus linguistics 26
4.2.1 Two-way test to compare third person singular reference in English and Japanese texts 26
4.2.2 Use of the chi-square test to show if data is normally distributed 27
4.2.3 Use of the chi-square test with word frequency lists to measure similarity between corpora 28
4.3 Correlation 29
4.3.1 The Pearson product–moment correlation coefficient 29
4.3.2 Use of Pearson's product–moment correlation coefficient in corpus linguistics 31
4.3.3 Spearman's rank correlation coefficient 32
4.3.4 Use of Spearman's rank correlation coefficient in corpus linguistics 32
4.4 Regression 33
4.4.1 Use of regression in corpus linguistics 34
4.4.2 Multiple regression 35
4.4.3 Use of multiple regression in corpus linguistics 36
5 Loglinear modelling 37
5.1 Overview 37
5.2 Iterative proportional scaling 39
5.3 Selection of the best model 42
5.4 Example of the use of loglinear analysis in corpus linguistics: gradience in the use of the genitive 43
6 Bayesian statistics 47
6.1 Use of Bayesian statistics in corpus linguistics 48
7 Summary 49
8 Exercises 50
9 Further reading 51
Notes 51

2 Information theory 53
1 Introduction 53
2 Basic formal concepts and terms 54
2.1 Language models 54

2.2 Shannon's theory of communication 55
2.3 Comparison between signal processing, text and speech 56
2.4 Information and entropy 58
2.5 Optimal codes 60
2.6 Redundancy 61
2.7 Mutual information 63
2.8 Stochastic processes: a series of approximations to natural language 65
2.9 Discrete Markov processes 67
2.10 Hidden Markov models 68
2.10.1 Evaluation: the forward–backward algorithm 69
2.10.2 Estimation: the Viterbi algorithm 72
2.10.3 Training: the Baum–Welch algorithm 73
2.11 Perplexity 75
2.12 Use of perplexity in language modelling 76
3 Probabilistic versus rule-based models of language 76
4 Uses of information theory in natural language processing: case studies 80
4.1 Part-of-speech taggers 80
4.1.1 The CLAWS word-tagging system: a corpus annotation tool 81
4.1.2 The Cutting tagger 84
4.1.3 Other applications of the Markov model in natural language processing 85
4.2 Use of redundancy in corpus linguistics 85
4.3 An experimental method for determining the amount of information and redundancy: Shannon's game 85
4.4 Information theory and word segmentation 86
4.5 Information theory and secrecy systems 87
4.6 Information theory and stylistics 88
4.7 Use of mutual information in corpus linguistics 89
5 The relationship between entropy, chi-square and the multinomial theorem 90
6 Summary 91
7 Exercises 93
8 Further reading 93
Notes 94

3 Clustering 95
1 Introduction 95
2 Reducing the dimensionality of multivariate data 96
2.1 Principal components analysis 97
2.2 Conversion of loadings to weights 101

2.3 Deciding on the number of principal components and their interpretation 101
2.4 The covariance matrix and the correlation matrix 102
2.5 Principal component analysis and spoken data: Horvath's study of vowel variation in Australian English 102
2.6 Factor analysis 105
2.7 Factor analysis in corpus linguistics: the work of Biber and Finegan 105
2.8 Factor analysis and a corpus of texts from the French Revolution 107
2.9 Mapping techniques 108
2.9.1 Principal coordinates analysis 108
2.9.2 Multi-dimensional scaling 109
3 Cluster analysis 110
3.1 Document clustering: the measurement of interdocument similarity 111
3.2 Distance coefficients 111
3.3 Association coefficients 112
3.4 Probabilistic similarity coefficients 114
3.5 Correlation coefficients 114
3.6 Non-hierarchic clustering 115
3.7 Hierarchic clustering methods 116
3.8 Types of hierarchical agglomerative clustering techniques 117
3.9 The validity of document clustering 120
4 Approximate string matching techniques: clustering of words according to lexical similarity 120
4.1 Equivalence and similarity 121
4.2 Term truncation 122
4.3 Suffix and prefix deletion and substitution rules 122
4.4 Clustering of words according to the constituent *n*-grams 125
4.5 Longest common subsequence and longest common substring 126
4.6 Dynamic programming 127
4.7 SPEEDCOP 129
4.8 Soundex 130
4.9 Corpus-based applications of approximate string matching 131
4.9.1 Typical spelling errors 131
4.9.2 Use of context to disambiguate real-word errors 133
4.9.3 Automatic sentence alignment in parallel corpora 135
4.9.4 Use of approximate string matching techniques to identify historical variants 137
5 Clustering of terms according to semantic similarity 138
5.1 Zernik's method of tagging word sense in a corpus 138
5.2 Phillips's experiment on the lexical structure of science texts 139

5.3 Morphological analysis of chemical and medical terms 141
5.4 Automatic thesaurus construction 143
6 Clustering of documents according to sublanguage using the perplexity measure 144
7 Clustering of dialects in Gaelic 145
8 Summary 146
9 Exercises 147
10 Further reading 148
Notes 148

4 Concordancing, collocations and dictionaries 149
1 Introduction 149
2 The concordance: introduction 150
2.1 Determining the context of a concordance 152
2.2 Text selection and preparation for concordancing 153
2.3 Modes of sorting the output from a concordance 155
2.4 Simple concordance packages 156
2.4.1 COCOA 156
2.4.2 The Oxford Concordance Program 157
2.4.3 EYEBALL 157
2.4.4 WordCruncher 158
3 Collocations: introduction 158
3.1 Syntactic criteria for collocability 161
3.2 Measures of collocation strength 162
3.2.1 Berry-Rogghe's *z*-score calculation 163
3.2.2 Collocational strength in French political tracts 166
3.2.3 The use of combinatorics to determine collocational significance 167
3.2.4 Daille's approach to monolingual terminology extraction 169
3.2.5 Mutual information for the extraction of bilingual word pairs 174
3.2.6 Pattern affinities and the extraction of bilingual terminology 175
3.2.7 Kay and Röscheisen's text–translation alignment algorithm 177
3.2.8 Production of probabilistic dictionaries from pre-aligned corpora 179
3.2.9 Collocations involving more than two words: the cost criterion 180
3.2.10 Factor analysis to show co-occurrence patterns among collocations 180
3.2.11 Mutual information for word segmentation 182

3.2.12 Smadja's XTRACT program for the extraction of collocations 182

3.2.13 Extraction of multi-word collocation units by *Champollion* 184

3.2.14 Use of a bilingual corpus to disambiguate word sense 186

3.2.15 *G*-square or log likelihood in the development of a 'phrasicon' 189

3.2.16 Dispersion 189

3.2.17 Hayashi's quantification method 192

4 Concordancing tools using more advanced statistics 193

4.1 WordSmith 193

4.2 CobuildDirect 194

4.3 ICAME's Lexa 194

4.4 TACT 195

4.5 The Hua Xia concordancer for Chinese text 195

5 Summary 195

6 Exercises 196

7 Further reading 197

Notes 197

5 Literary detective work 199

1 Introduction 199

2 The statistical analysis of writing style 200

2.1 Morton's criteria for the computation of style 200

2.2 Early work on the statistical study of literary style 202

2.3 The authorship of *The Imitation of Christ* 203

2.4 The authorship of the New Testament Epistles 205

2.5 Ellegård's work on the Junius letters 207

2.6 The Federalist papers 208

2.6.1 The Bayesian approach of Mosteller and Wallace 208

2.6.2 Biological analogy and the Federalist papers: recent developments 212

2.7 Kenny's work on the Aristotelean *Ethics* 214

2.8 A stylometric analysis of Mormon scripture 214

2.9 Studies of Swift and his contemporaries 214

2.10 Was *And Quiet Flows the Don* written by Mikhail Sholokhov? 218

2.11 The authenticity of the Baligant episode in the *Chanson de Roland* 218

2.12 The chronology of Isocrates 219

2.13 Studies of Shakespeare and his contemporaries 221

2.14 Comparison of Kierkegaard's pseudonyms 223

2.15 Dogmatic and tentative alternatives 224

2.16 Leighton's study of 17th-century German sonnets 225

2.17 Forensic stylometry 226
2.18 The use of a syntactically annotated corpus in stylometry 227
3 Studies of linguistic relationship 230
3.1 Ellegård's method for the statistical measurement of linguistic relationship 230
3.2 Language divergence and estimated word retention rate 232
4 Decipherment and translation 236
4.1 The decipherment of Linear B 236
4.2 The Hackness cross cryptic inscriptions 240
4.3 A statistical approach to machine translation 243
5 Summary 246
6 Exercises 247
7 Further reading 248
Note 248

Glossary 249
Appendices 258
Appendix 1 The normal distribution 258
Appendix 2 The *t* distribution 260
Appendix 3 The Mann–Whitney *U* test 261
Appendix 4 The sign test 262
Appendix 5 The Wilcoxon signed ranks test 263
Appendix 6 The *F* distribution 264
Appendix 7 The chi-square distribution 266
Appendix 8 The Pearson product–moment correlation coefficient 267
Appendix 9 The Spearman rank correlation coefficient 268
Answers to exercises 269
Bibliography 271
Index 283