Contents

List of Figures xiii

List of Tables xvii

Preface xix

Acknowledgements xxi

Introduction 1

1 Corpus Linguistics: Basic Principles 5

Introduction 5

1. Theory, approach, methods: Corpus linguistics as a research field 5

2. Key issues in corpus linguistics 7

2.1 Authenticity 9

2.2 Representativeness 10

2.3 Balance and sampling 12

2.4 Size 14

2.5 Types of corpora 15

3. *Corpus, concordance, collocation*: Tools and analysis 16

3.1 Corpus creation 16

3.2 Corpus analysis 18

3.2.1 Wordlists and keywords 19

3.2.2 Concordances 23

3.3 Collocation and basic statistics 25

3.4 Colligation and semantic associations 29

Conclusion 31

Study questions and activities 32

Suggestions for further reading 33

2 The Body and the Web: An Introduction to the Web as Corpus 35

Introduction 35

1. Corpus linguistics and the web 35

2. The web as corpus: A 'body' of texts? 39

3. The corpus and the web: Key issues 41

3.1 Authenticity 42

3.2 Representativeness 43

3.3 Size 45

3.4 Composition 49

3.4.1 Medium 51

3.4.2 Language 52

3.4.3 Topics 57

3.4.4 Registers, (web) genres, and text types 61

3.5 Copyright 63

4. From 'body' to 'web': New issues 65

4.1 Dynamism 66

4.2 Reproducibility 68

4.3 Relevance and reliability 69

Conclusion 70

Study questions and activities 71

Suggestions for further reading 71

3 Challenging Anarchy: Web Search from a Corpus Perspective 73

Introduction 73

1. The corpus and the search 73

2. Search engine basics: Crawling, indexing, searching, ranking 77

3. Google and the others: An overview of commercial search engines 79

4. Challenging anarchy: Mastering advanced web search 83

4.1 An overview of web search options 84

4.2 Limits and potentials of 'webidence' 87

4.3 Phrase search and collocation 88

4.4 Phraseology and patterns 91

4.5 Provenance, site, domain and more: Searching subsections of the web 93

4.6 Testing translation candidates 96

5. Query complexity: Web search from a corpus perspective 101

Conclusion 102

Study questions and activities 102
Suggestions for further reading 103

4 Beyond Ordinary Search Engines: Concordancing the Web 105

Introduction 105
1. Beyond ordinary search engines: Concordancing the web 105
2. WebCorp *Live* 106
3. Concordancing the web in the foreign language classroom 113
3.1 Exploring collocation: The case of *scenery* 113
3.2 Investigating neologisms and phrasal creativity 115
4. Beyond web concordancing tools: The web as/for corpus 119
5. Towards a linguist's search engine: The case of WebCorp*LSE* 121
5.1 The Web as/for Corpus: From WebCorp *Live* to WebCorp*LSE* 122
5.2 Using WebCorp*LSE* to explore contemporary English 125
5.2.1 Synchronic English Web Corpus and Diachronic English Web Corpus 126
5.2.2 Birmingham Blog Corpus 131
Conclusion 134
Study questions and activities 134
Suggestions for further reading 136

5 Building and Using Comparable Web Corpora: Tools and Methods 137

Introduction 137
1. *Building DIY web corpora* 137
2. From words to corpus: The 'bootstrap' process 140
2.1 Compiling a domain-specific corpus with BootCaT 140
2.2 Compiling specialized corpora with WebBootCaT 147
3. Building and using comparable web corpora for translation practice 154
Conclusion 158
Study questions and activities 159
Suggestions for further reading 161

6 Sketches of Language and Culture from Large Web Corpora 163

Introduction 163
1. From web as corpus to corpus as web: Introducing large general purpose web corpora 163

2. Mega–corpus, mini–Web: The case of ukWaC 167
2.1 Selecting 'seed URLs' and crawling 167
2.2 Post-crawl cleaning and annotation 169
3. Exploring large web corpora: Tools and resources 171
3.1 The Sketch Engine: From concordance lines to word sketches 172
3.2 The Sketch Difference function 179
4. Case study: Sketches of *culture* from the BNC to the web 182
4.1 A very difficult word... 183
4.2 Sketches of *culture* from the British National Corpus 184
4.3 Sketches of *culture* from UkWaC 187
4.3.1 *Culture* as object 188
4.3.2 *Culture* as subject 193
4.3.3 Modifiers of culture 195
4.3.4 *Culture* as modifier 196
4.3.5 The pattern *culture* and/or NOUN 198
4.4 *A culture of*: The changing face of culture in contemporary society 198
Conclusion 202
Study questions and activities 203
Suggestions for further reading 203

7 From Download to Upload: The Web as Corpus in the Web 2.0 Era 205

Introduction 205
1. From download to upload: Web users as *prosumers* 205
2. Web 2.0 as corpus. The case of Wikipedia as a multilingual corpus 207
3. The corpus in the cloud? The challenges that lie ahead 208
Suggestions for further reading 210

Conclusion 211

References 215
Index 229