

Contents

Preface	vii
Acknowledgments	ix
1 The need for corpus data	1
1.1 Arguments against corpus data	2
1.1.1 Corpus data as usage data	3
1.1.2 The incompleteness of corpora	5
1.1.3 The absence of meaning in corpora	7
1.2 Intuition	8
1.2.1 Intuition as performance	11
1.2.2 The incompleteness of intuition	11
1.2.3 Intuitions about form and meaning	13
1.3 Intuition data vs. corpus data	15
1.4 Corpus data in other sub-disciplines of linguistics	17
2 What is corpus linguistics?	21
2.1 The linguistic corpus	22
2.1.1 Authenticity	23
2.1.2 Representativeness	28
2.1.3 Size	37
2.1.4 Annotations	38
2.2 Towards a definition of corpus linguistics	46
2.3 Corpus linguistics as a scientific method	56
3 Corpus linguistics as a scientific method	61
3.1 The scientific hypothesis	61
3.1.1 Stating hypotheses	62
3.1.2 Testing hypotheses: From counterexamples to probabilities	68
3.2 Operationalization	77
3.2.1 Operational definitions	77
3.2.2 Examples of operationalization in corpus linguistics . .	81
3.2.2.1 Parts of speech	84

Contents

3.2.2.2	Length	90
3.2.2.3	Discourse status	93
3.2.2.4	Word senses	96
3.2.2.5	Animacy	98
3.2.2.6	Interim summary	100
3.3	Hypotheses in context: The research cycle	100
4	Data retrieval and annotation	105
4.1	Retrieval	106
4.1.1	Corpus queries	106
4.1.2	Precision and recall	111
4.1.3	Manual, semi-manual and automatic searches	116
4.2	Annotating	120
4.2.1	Annotating as interpretation	121
4.2.2	Annotation schemes	122
4.2.3	The reliability of annotation schemes	127
4.2.4	Reproducibility	133
4.2.5	Data storage	136
5	Quantifying research questions	141
5.1	Types of data	141
5.1.1	Nominal data	143
5.1.2	Ordinal data	145
5.1.3	Cardinal data	146
5.1.4	Interim summary	148
5.2	Descriptive statistics for nominal data	148
5.2.1	Percentages	151
5.2.2	Observed and expected frequencies	154
5.3	Descriptive statistics for ordinal data	157
5.3.1	Medians	161
5.3.2	Frequency lists and mode	162
5.4	Descriptive statistics for cardinal data	163
5.4.1	Means	164
5.5	Summary	166
6	Significance testing	167
6.1	Statistical hypothesis testing	167
6.2	Probabilities and significance testing	170

6.3	Nominal data: The chi-square test	177
6.3.1	Two-by-two designs	177
6.3.2	One-by- <i>n</i> designs	184
6.4	Ordinal data: The Mann-Whitney <i>U</i> -test	187
6.5	Inferential statistics for cardinal data	191
6.5.1	Welch's <i>t</i> -test	192
6.5.2	Normal distribution requirement	196
6.6	Complex research designs	199
6.6.1	Variables with more than two values	199
6.6.2	Designs with more than two variables	203
6.6.2.1	A danger of bivariate designs	205
6.6.2.2	Configural frequency analysis	208
7	Collocation	215
7.1	Collocates	215
7.1.1	Collocation as a quantitative phenomenon	217
7.1.2	Methodological issues in collocation research	221
7.1.3	Effect sizes for collocations	224
7.1.3.1	Chi-square	225
7.1.3.2	Mutual Information	226
7.1.3.3	The log-likelihood ratio test	227
7.1.3.4	Minimum Sensitivity	228
7.1.3.5	Fisher's exact test	228
7.1.3.6	A comparison of association measures	229
7.2	Case studies	233
7.2.1	Collocation for its own sake	233
7.2.1.1	Case study: Degree adverbs	234
7.2.2	Lexical relations	235
7.2.2.1	Case study: Near synonyms	235
7.2.2.2	Case study: Antonymy	240
7.2.3	Semantic prosody	244
7.2.3.1	Case study: True feelings	245
7.2.3.2	Case study: The verb <i>cause</i>	249
7.2.4	Cultural analysis	254
7.2.4.1	Case study: Small boys, little girls	255
8	Grammar	261
8.1	Grammar in corpora	261

Contents

8.2 Case studies 263

8.2.1 Collocational frameworks and grammar patterns 263

8.2.1.1 Case study: [*a* __ *of*] 264

8.2.1.2 Case study: [*there* V_{link} *something* ADJ *about* NP] 268

8.2.2 Collostructional analysis 270

8.2.2.1 Case study: The ditransitive 270

8.2.2.2 Case study: Ditransitive and prepositional dative 272

8.2.2.3 Case study: Negative evidence 272

8.2.3 Words and their grammatical properties 276

8.2.3.1 Case study: Complementation of *begin* and *start* 276

8.2.4 Grammar and context 282

8.2.4.1 Case study: Adjective order and frequency . . 283

8.2.4.2 Case study: Binomials and sonority 286

8.2.4.3 Case study: Horror aequi 289

8.2.4.4 Case study: Synthetic and analytic comparatives and persistence 292

8.2.5 Variation and change 295

8.2.5.1 Case study: Sex differences in the use of tag questions 295

8.2.5.2 Case study: Language change 298

8.2.5.3 Case study: Grammaticalization 300

8.2.6 Grammar and cultural analysis 303

8.2.6.1 Case study: He said, she said 303

8.2.7 Grammar and counterexamples 305

8.2.7.1 Case study: *To-* vs. *that-*complements 305

9 Morphology 309

9.1 Quantifying morphological phenomena 310

9.1.1 Counting morphemes: Types, tokens and hapax legomena 310

9.1.1.1 Token frequency 312

9.1.1.2 Type frequency 315

9.1.1.3 Hapax legomena 316

9.1.2 Statistical evaluation 318

9.2 Case studies 325

9.2.1 Morphemes and stems 325

9.2.1.1 Case study: Phonological constraints on *-ify* . 326

9.2.1.2 Case study: Semantic differences between *-ic* and *-ical* 330

9.2.1.3	Case study: Phonological differences between -ic and -ical	338
9.2.1.4	Case study: Affix combinations	340
9.2.2	Morphemes and demographic variables	343
9.2.2.1	Case study: Productivity and genre	343
9.2.2.2	Case study: Productivity and speaker sex	348
10	Text	353
10.1	Keyword analysis	353
10.2	Case studies	358
10.2.1	Language variety	358
10.2.1.1	Case study: Keywords in scientific writing	358
10.2.1.2	Case study: [a + ___ + of] in Scientific English	359
10.2.2	Comparing speech communities	361
10.2.2.1	Case study: British vs. American culture	367
10.2.2.2	Case study: “African” keywords	370
10.2.3	Co-occurrence of lexical items and demographic categories	372
10.2.3.1	Case study: A deductive approach to sex dif- ferences	372
10.2.3.2	Case study: An inductive approach to sex dif- ferences	378
10.2.4	Ideology	380
10.2.4.1	Case study: Political ideologies	380
10.2.4.2	Case study: The importance of men and women	384
10.2.5	Time periods	387
10.2.5.1	Case study: Verbs in the <i>going-to</i> future	387
10.2.5.2	Case study: Culture across time	392
11	Metaphor	397
11.1	Studying metaphor in corpora	397
11.2	Case studies	398
11.2.1	Source domains	398
11.2.1.1	Case study: Lexical relations and metaphorical mapping	399
11.2.1.2	Case study: Word forms in metaphorical map- pings	402
11.2.1.3	Case study: The impact of metaphorical ex- pressions	406

Contents

11.2.2	Target domains	410
11.2.2.1	Case study: Happiness across cultures	412
11.2.2.2	Case study: Intensity of emotions	417
11.2.3	Metaphor and text	422
11.2.3.1	Case study: Identifying potential source domains	422
11.2.3.2	Case study: Metaphoricity signals	425
11.2.3.3	Case study: Metaphor and ideology	431
11.2.4	Metonymy	434
11.2.4.1	Case study: Subjects of the verb <i>bomb</i>	434
12	Epilogue	437
13	Study notes	441
14	Statistical tables	447
14.1	Critical values for the chi-square test	447
14.2	Chi-square table for multiple tests with one degree of freedom .	448
14.3	Critical values for the Mann-Whitney-Text	449
14.4	Critical values for Welch's <i>t</i> -Test	451
	References	453
	Index	477
	Name index	477
	Subject index	481