

Corpus-Based Language Studies

An advanced resource book

Tony McEnergy, Richard Xiao and Yukio Tono

Contents

Series editors' preface	xv
Preface	xvii
Acknowledgements	xix

SECTION A: INTRODUCTION **1**

Unit A1 Corpus linguistics: the basics	3
A1.1 Introduction	3
A1.2 Corpus linguistics: past and present	3
A1.3 What is a corpus?	4
A1.4 Why use computers to study language?	5
A1.5 The corpus-based approach vs. the intuition-based approach	6
A1.6 Corpus linguistics: a methodology or a theory?	7
A1.7 Corpus-based vs. corpus-driven approaches	8
Summary	11
Looking ahead	12
Unit A2 Representativeness, balance and sampling	13
A2.1 Introduction	13
A2.2 What does representativeness mean in corpus linguistics?	13
A2.3 The representativeness of general and specialized corpora	15
A2.4 Balance	16
A2.5 Sampling	19
Summary	21
Looking ahead	21
Unit A3 Corpus mark-up	22
A3.1 Introduction	22
A3.2 The rationale for corpus mark-up	22
A3.3 Corpus mark-up schemes	23
A3.4 Character encoding	27
Summary	28
Looking ahead	28
Unit A4 Corpus annotation	29
A4.1 Introduction	29
A4.2 Corpus annotation = added value	30
A4.3 How is corpus annotation achieved?	33
A4.4 Types of corpus annotation	33
A4.5 Embedded vs. standalone annotation	44

Summary	44
Looking ahead	45
Unit A5 Multilingual corpora	46
A5.1 Introduction	46
A5.2 Multilingual corpora: terminological issues	47
A5.3 Corpus alignment	50
Summary	51
Looking ahead	51
Unit A6 Making statistical claims	52
A6.1 Introduction	52
A6.2 Raw frequency and normalized frequency	52
A6.3 Descriptive and inferential statistics	53
A6.4 Tests of statistical significance	53
A6.5 Tests for significant collocations	56
Summary	57
Looking ahead	57
Unit A7 Using available corpora	59
A7.1 Introduction	59
A7.2 General corpora	59
A7.3 Specialized corpora	60
A7.4 Written corpora	61
A7.5 Spoken corpora	62
A7.6 Synchronic corpora	64
A7.7 Diachronic corpora	65
A7.8 Learner corpora	65
A7.9 Monitor corpora	67
Summary	69
Looking ahead	70
Unit A8 Going solo: DIY corpora	71
A8.1 Introduction	71
A8.2 Corpus size	71
A8.3 Balance and representativeness	73
A8.4 Data capture	73
A8.5 Corpus mark-up	74
A8.6 Corpus annotation	75
A8.7 Character encoding	76
Summary	76
Looking ahead	76
Unit A9 Copyright	77
A9.1 Introduction	77
A9.2 Coping with copyright: warning and advice	77
Summary	78
Looking ahead	79
Unit A10 Corpora and applied linguistics	80
A10.1 Introduction	80
A10.2 Lexicographic and lexical studies	80

A10.3	Grammatical studies	85
A10.4	Register variation and genre analysis	87
A10.5	Dialect distinction and language variety	90
A10.6	Contrastive and translation studies	91
A10.7	Diachronic study and language change	96
A10.8	Language learning and teaching	97
A10.9	Semantics	103
A10.10	Pragmatics	104
A10.11	Sociolinguistics	108
A10.12	Discourse analysis	111
A10.13	Stylistics and literary studies	113
A10.14	Forensic linguistics	116
A10.15	What corpora cannot tell us	120
	Summary	121
	Looking ahead	122
SECTION B: EXTENSION		123
Unit B1	Corpus representativeness and balance	125
B1.1	Introduction	125
B1.2	Biber (1993)	125
B1.3	Atkins, Clear and Ostler (1992)	128
	Summary	130
	Looking ahead	130
Unit B2	Objections to corpora: an ongoing debate	131
B2.1	Introduction	131
B2.2	Widdowson (2000)	131
B2.3	Stubbs (2001b)	135
B2.4	Widdowson (1991) vs. Sinclair (1991b): a summary	140
	Summary	144
Unit B3	Lexical and grammatical studies	145
B3.1	Introduction	145
B3.2	Krishnamurthy (2000)	145
B3.3	Partington (2004)	148
B3.4	Carter and McCarthy (1999)	152
B3.5	Kreyer (2003)	155
	Summary	159
	Looking ahead	159
Unit B4	Language variation studies	160
B4.1	Introduction	160
B4.2	Biber (1995a)	160
B4.3	Hyland (1999)	165
B4.4	Lehmann (2002)	169
B4.5	Kachru (2003)	174
	Summary	177
	Looking ahead	177
Unit B5	Contrastive and diachronic studies	178
B5.1	Introduction	178

B5.2	Altenberg and Granger (2002)	178
B5.3	McEnery, Xiao and Mo (2003)	181
B5.4	Kilpiö (1997)	185
B5.5	Mair, Hundt, Leech and Smith (2002)	190
	Summary	194
	Looking ahead	194
Unit B6	Language teaching and learning	195
B6.1	Introduction	195
B6.2	Gavioli and Aston (2001)	195
B6.3	Thurston and Candlin (1998)	198
B6.4	Conrad (1999)	201
	Summary	202
	Looking ahead	203
SECTION C: EXPLORATION		205
Unit C1	Collocation and pedagogical lexicography Case study 1	208
C1.1	Introduction	208
C1.2	Collocation information	210
C1.3	Using corpus data for improving a dictionary entry	220
	Summary	225
	Further study	225
Unit C2	HELP or HELP to: what do corpora have to say? Case study 2	227
C2.1	Introduction	227
C2.2	Concordancing	228
C2.3	Language variety	235
C2.4	Language change	239
C2.5	An intervening NP	240
C2.6	The infinite marker preceding HELP	241
C2.7	The passive construction	245
	Summary	246
	Further study	246
Unit C3	L2 acquisition of grammatical morphemes Case study 3	247
C3.1	Introduction	247
C3.2	Morpheme studies: a short review	249
C3.3	The Longman Learners' Corpus	250
C3.4	Problem-oriented corpus annotation	251
C3.5	Discussion	260
	Summary	263
	Further study	263
Unit C4	Swearing in modern British English Case study 4	264
C4.1	Introduction	264
C4.2	Spoken vs. written register	265
C4.3	Variations within spoken English	269
C4.4	Variations within written English	279
	Summary	285
	Further study	286

Unit C5	Conversation and speech in American English Case study 5	287
C5.1	Introduction	287
C5.2	Salient linguistic features	288
C5.3	Basic statistical data from the corpus	293
C5.4	The dimension scores of three genres	303
C5.5	The keyword approach to genre analysis	308
	Summary	319
	Further study	320
Unit C6	Domains, text types, aspect marking and English–Chinese translation Case study 6	321
C6.1	Introduction	321
C6.2	The corpus data	323
C6.3	Translation of aspect marking	324
C6.4	Translation and aspect marking	336
C6.5	Domain and aspect marking	338
C6.6	Text type and aspect marking	340
	Summary	341
	Further study	343
	Glossary	344
	Bibliography	352
	Appendix of useful Internet links	379
	Index	381