

UNDERSTANDING LANGUAGE SERIES

Danielle Barth and
Stefan Schnell

Understanding
Corpus Linguistics



สำนักหอสมุด มหาวิทยาลัยเชียงใหม่

b16724 80x

012585427

j22706823



Understanding

Corpus Linguistics

Danielle Barth and
Stefan Schnell

 **Routledge**
Taylor & Francis Group
LONDON AND NEW YORK

Contents

Acknowledgements	ix
1 Introduction	1
1.1 What is corpus linguistics?	1
1.1.1 The basic idea of corpus linguistics	1
1.1.2 Corpus linguistics in contrast to other approaches	2
1.1.3 Corpus linguistics and usage-oriented linguistics	4
1.2 Rundown of this book	5
2 Basic concepts in corpus linguistics	7
2.1 What is a corpus and what is it good for?	7
2.2 Definition and characteristics of texts in corpus linguistics	8
2.2.1 What is a text in corpus linguistics?	8
2.2.2 What is a word in a text?	10
2.2.3 Types and tokens	11
2.2.4 Textual and contextual properties of texts	13
2.3 Corpora as samples of language use	17
2.4 Conclusion	18
3 Corpus composition and corpus types	19
3.1 Corpus content and representativeness	19
3.1.1 Corpus size	19
3.1.2 Corpus composition	22
3.1.3 Authenticity, routines, and spontaneity	25
3.1.4 Representativeness	27
3.1.5 Saturation	28
3.1.6 Text varieties: register, genre, and style	29
3.2 Linked data	32
3.2.1 Raw data and primary data	32
3.2.2 Metadata	33
3.3 Corpus types	34
3.3.1 Corpus types defined by situational features	34
4 Levels of linguistic representation in corpus-linguistic research	41
4.1 Linguistic structures and their variants	41
4.1.1 Language use and contextualisation	41
4.1.2 Choice of variants: possible and probable structures	42
4.2 Structural levels of variation	44
4.2.1 Lexical semantics	44

4.2.2	Phonetics	49
4.2.3	Phonology	52
4.2.4	Morphology	53
4.2.5	Syntax	57
4.2.6	Discourse	59
4.3	Sign and gesture	63
4.4	Conclusion	67
5	Corpus queries	68
5.1	Getting started	68
5.2	Corpus queries	68
5.3	Frequency lists	71
5.3.1	Keywords	73
5.4	Graphical frequency descriptions	73
5.4.1	Frequency plots	73
5.4.2	Dispersion plots	74
5.5	Zipfian distribution	75
5.6	Collocations and bigrams	77
5.7	Association measures for bigrams	79
5.8	Concordances and keywords in context	82
5.9	Trigrams and n-grams	83
5.10	Colligations and grammatical categories	83
5.11	Regular expressions and specialised query languages	85
5.11.1	Regex with unicode	88
5.12	Workflow	89
5.13	Conclusion	89
5.14	Answers for exercise 5.6 on regexes	90
6	Corpus building	91
6.1	Steps for an idealised corpus	91
6.1.1	Corpus types and goals	91
6.1.2	Identification, selection, and evaluation	94
6.2	Corpus building in context	96
6.2.1	Whole-population corpora	97
6.2.2	Availability of texts: copyright and privacy	97
6.2.3	Technical, methodological, and other considerations	98
6.3	Pre-processing: transcription, translation, and digitisation of text	100
6.3.1	Spoken data transcription	100
6.3.2	Linking transcriptions to raw data	102
6.3.3	Rendering corpus text	104
6.3.4	Translation	104
6.3.5	Choice of metalanguage	105
6.4	Data formats	106
6.5	Including metadata	107
6.6	Publication of the corpus	108
6.7	Conclusion	108

7	Corpus annotation	110
7.1	Corpus annotations and research agendas	110
7.1.1	Standards of corpus annotation	111
7.2	Types of annotation and annotation schemes	111
7.2.1	Annotation for phonetics and prosody	112
7.2.2	Morphological annotation	113
7.2.3	Parts-of-speech tagging	115
7.2.4	Syntactic annotation	119
7.2.5	Semantic annotation	123
7.2.6	Discourse and reference annotation	123
7.3	Corpus annotation in linguistic typology	125
7.3.1	UDs: Universal dependencies	126
7.3.2	SCOPIC: Social cognition parallax interview corpus	127
7.3.3	Multi-CAST: Multilingual corpus of annotated spoken texts	129
7.4	Conclusion	134
8	Statistical description and analysis	136
8.1	Introduction	136
8.2	Basics	137
8.2.1	Sampling	137
8.2.2	Dependent and independent variables	137
8.2.3	Distributions	138
8.2.4	Range and spread	140
8.3	Common univariate tests	142
8.3.1	Null hypothesis	142
8.3.2	Chi-squared test	142
8.3.3	Correlations	145
8.4	Multivariate predictive approaches	146
8.4.1	Mixed-effects multiple regression	149
8.4.2	Recursive partitioning	156
8.4.3	Clustering methods	161
8.5	Making statistical claims	162
8.6	Conclusion	163
9	Corpora in sociolinguistics	164
9.1	A brief introduction to sociolinguistics	164
9.1.1	Overlap in studies of variation	164
9.1.2	Variables in sociolinguistics	166
9.2	Dialect and regional variation	167
9.2.1	Cross-dialectal variation in large corpora	168
9.2.2	Dialectal variation in corpora of sociolinguistic interviews	168
9.2.3	Corpus-based dialectometry	170
9.3	Social factors of variation and change	171
9.3.1	Ethnicity	171
9.3.2	Sex and gender	173
9.3.3	Interaction of constraints	175

9.4	Variation and language change	177
9.5	Conclusion	180
10	Corpus linguistics and language documentation	182
10.1	Language documentation: Capturing the diversity of human languages	182
10.1.1	Defining language documentation	183
10.1.2	Language documentation in practice	184
10.1.3	What are “linguistic practices”?	185
10.2	Lasting records of language use and their uses in corpus linguistics	186
10.2.1	Language documentation and data types generated	186
10.2.2	Multiple purposes of language documentation	187
10.3	From Collection to corpus: corpus building from language documentation	187
10.3.1	Corpora versus language documentations	190
10.4	Research questions appropriate for small corpora	190
10.4.1	Function words	191
10.4.2	Grammatical marking	192
10.4.3	Classes of words	193
10.4.4	Constructions (syntactic variants)	193
10.4.5	Zeros	193
10.4.6	Phonetic information	194
10.4.7	What we can't study	195
10.4.8	Advantages of small corpora	195
10.5	Conclusion	196
11	Corpus-based typology	197
11.1	Introduction: Typological questions and linguistic corpora	197
11.2	Universals and diversity in language use	199
11.2.1	What is universal in language use?	199
11.2.2	Universal constraints on diverse patterns of language use	203
11.2.3	How does language use differ across languages and cultures?	206
11.3	Issues in corpus development and cross-corpus typological research design	212
11.3.1	Different types of multilingual corpora	212
11.3.2	Corpus property biases in CBT research	213
11.3.3	Written and LOL (‘literate’, ‘official’, ‘lots of speakers’) biases in corpus-based typology	214
11.3.4	Corpus design versus bootstrapping for specific research questions	214
11.4	Conclusion	215
	References	217
	Index	233