

Contents

<i>List of Figures</i>	page viii
<i>List of Tables</i>	x
<i>Acknowledgments</i>	xii
1 Introduction	1
1.1 Introduction	1
1.2 Our Operational Definition of a “Corpus”	7
1.2.1 A Sample of <u>Texts</u>	7
1.2.2 A Corpus Is <u>Large</u>	9
1.2.3 A <u>Principled</u> Sample <u>Designed</u> to <u>Represent</u> a Domain	10
1.3 A Preliminary Definition of Representativeness in Corpus Linguistics	11
1.4 Target Audiences for <i>DELC</i>	13
1.5 Outline and Key Features of the Book	15
Chapter 1 Exercises and Discussion Points	18
2 Approaches to Representativeness in Previous Corpus Linguistic Research	28
2.1 What Is the Statistical Meaning of REPRESENTATIVENESS?	28
2.2 A Survey of Previous Conceptualizations of Representativeness in Corpus Linguistics	30
2.2.1 Representativeness = “GENERAL ACCLAIM FOR DATA”	30
2.2.2 Representativeness = “ABSENCE OF SELECTIVE FORCES”	31
2.2.3 Representativeness = “TYPICAL OR IDEAL CASES”	32
2.2.4 Representativeness = “MINIATURE OF THE POPULATION”	33
2.2.5 Representativeness = “COVERAGE OF THE POPULATION’S HETEROGENEITY”	34
2.2.6 Representative = “PERMITTING GOOD ESTIMATION”	35
2.2.7 Representativeness = “DESIGNED FOR A PARTICULAR PURPOSE”	36
2.2.8 A VERY LARGE CORPUS IS A DE FACTO REPRESENTATIVE CORPUS	36
2.2.9 A BALANCED CORPUS IS A REPRESENTATIVE CORPUS	37
2.2.10 A REPRESENTATIVE CORPUS IS NEVER POSSIBLE	39
2.3 Chapter Summary	41
Chapter 2 Exercises and Discussion Points	43

3	Corpus Representativeness: A Conceptual and Methodological Framework	52
3.1	Overview and Definitions	52
3.2	Linguistic Parameter Estimation – The Ultimate Objective of Corpus Linguistic Analysis	56
3.3	Linguistic Research Goals	57
3.4	Domain Considerations	58
3.5	Distribution Considerations	60
3.6	Corpus Representativeness Requires both Domain and Distribution Considerations	61
3.7	Representativeness as a Continuous Construct	62
3.8	Chapter Summary	63
	Chapter 3 Exercises and Discussion Points	64
4	Domain Considerations	68
4.1	Introduction	68
4.2	Describing the Domain	71
4.2.1	Methods and Resources for Domain Description	81
4.2.2	Defining Domain Boundaries	86
4.2.3	Establishing Domain-Internal Categories	88
4.3	Operationalizing the Domain	91
4.3.1	Specifying Operational Domain Boundaries and Strata	93
4.3.2	Evaluation: Operational Domain → Domain	97
4.4	Sampling the Texts	98
4.4.1	Sampling Units and Sampling Designs	98
4.4.2	Stratification	100
4.4.3	Relative Sizes of the Strata	100
4.4.4	Randomness	103
4.4.5	Nonrandom Sampling Methods	103
4.4.6	Evaluation: Corpus → Operational Domain	104
4.5	Detailed Case Study: From Domain Analysis to Corpus Design in the AJRC	105
4.6	Conclusion	112
	Chapter 4 Exercises and Discussion Points	113
5	Distribution Considerations	122
5.1	Introduction	122
5.2	Linguistic Variables	123
5.3	Sample Size	124
5.3.1	Undersampling	125
5.3.2	Oversampling	128
5.4	Analyzing Sample Size and Precision for Linguistic Rates of Occurrence	129
5.4.1	Determining Required Sample Size for Creating a New Corpus	130
5.4.2	Determining Precision for an Existing Corpus	134
5.4.3	Common Misconceptions about Sample Size	136
5.5	Achieving Precise Analyses of Linguistic Types	138
5.5.1	Corpora That Contain As Many Different Words As Possible	139
5.5.2	Creating a Rank-Ordered List of Linguistic Types	142
	Chapter 5 Exercises and Discussion Points	152

6	The Influence of Domain and Distribution Considerations on Corpus Representativeness – Bringing It All Together	156
6.1	Corpus Representativeness and Linguistic Parameter Estimation	156
6.2	Experimentally Investigating Domain and Distribution Considerations As Predictors of Quantitative-Linguistic Accuracy	160
6.2.1	Methods for the Experiments	161
6.2.2	Results of the Experiments	166
	Chapter 6 Exercises and Discussion Points	174
7	Corpus Design and Representativeness in Practice	177
	WITH DANIEL KELLER	177
7.1	Introduction	177
7.2	Key Steps in Ensuring and Evaluating Corpus Representativeness	178
7.3	Designing and Creating Representative Corpora: Two Case Studies	182
7.3.1	Case Study 1: Designing and evaluating the representativeness of a Corpus of Yelp Restaurant Reviews	182
7.3.2	Case Study 2: Designing and evaluating the representativeness of a Corpus of YouTube Vlogs	191
7.3.3	Summary: Addressing Challenges in the Creation of New Corpora	199
7.4	Evaluating the Suitability of an Existing Corpus for a Particular Research Question: Academic Research Writing	201
7.5	Conclusion	216
	Chapter 7 Exercises and Discussion Points	218
	<i>Glossary</i>	220
	<i>Appendix A List of Example Stand-alone Corpus Description Articles</i>	224
	<i>Appendix B Survey of Corpus Design and Compilation Practices</i>	226
B1	Corpus Survey	226
B2	Synthesis and Commentary	258
B2.1	Corpus Description	258
B2.2	Domain Description	261
B2.3	Evaluation and Documentation	268
B3	Looking Ahead	270
	<i>References</i>	271
	<i>Index</i>	280