## Contents

<table>
<thead>
<tr>
<th>List of Figures</th>
<th>page viii</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>xii</td>
</tr>
</tbody>
</table>

1 Introduction 1

1.1 Introduction 1

1.2 Our Operational Definition of a “Corpus” 7

1.2.1 A Sample of Texts 7

1.2.2 A Corpus Is Large 9

1.2.3 A Principled Sample Designed to Represent a Domain 10

1.3 A Preliminary Definition of Representativeness in Corpus Linguistics 11

1.4 Target Audiences for DELC 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

© in this web service Cambridge University Press www.cambridge.org
vi Contents

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