Corpora are ubiquitous in linguistic research, yet to date, there has been no consensus on how to conceptualize corpus representativeness and collect corpus samples. This pioneering book bridges this gap by introducing a conceptual and methodological framework for corpus design and representativeness. Written by experts in the field, it shows how corpora can be designed and built in a way that is both optimally suited to specific research agendas and adequately representative of the types of language use in question. It considers questions such as what types of texts should be included in the corpus and how many texts are required – highlighting that the degree of representativeness rests on the dual pillars of domain considerations and distribution considerations. The authors introduce, explain, and illustrate all aspects of this corpus representativeness framework in a step-by-step fashion, using examples and activities to help readers develop practical skills in corpus design and evaluation.

Jesse Egbert is Associate Professor of Applied Linguistics at Northern Arizona University. He is a cofounding General Editor of *Register Studies*, and his recent books focus on online register variation (2018), methodological triangulation (2016, 2020), and corpus linguistics methods (2020).

Douglas Biber is Regents’ Professor of Applied Linguistics at Northern Arizona University. His previous books include *Register, Genre, and Style* (2009/19), the *Grammar of Spoken and Written English* (2021), and studies of register variation (1988, 1995, 2018).

Bethany Gray is Associate Professor of Applied Linguistics and Technology at Iowa State University. Her publications include monographs on academic research articles (2015) and historical change in writing (2016). She is a cofounding General Editor of *Register Studies*. 
Designing and Evaluating Language Corpora

A Practical Framework for Corpus Representativeness

Jesse Egbert
Northern Arizona University

Douglas Biber
Northern Arizona University

Bethany Gray
Iowa State University
Contents

List of Figures .......... viii
List of Tables .......... x
Acknowledgments ......... xi

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 HETERGENEITY” .......... 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
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
Contents

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 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

Glossary 220

Appendix A List of Example Stand-alone Corpus Description Articles 224

Appendix B Survey of Corpus Design and Compilation Practices 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

References 271

Index 280
# Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Key concepts included in the definition of a “corpus”</td>
<td>3</td>
</tr>
<tr>
<td>1.2</td>
<td>Distribution of research articles employing corpus linguistics in six subdisciplines</td>
<td>19</td>
</tr>
<tr>
<td>3.1</td>
<td>Graphical representation of our representativeness framework</td>
<td>54</td>
</tr>
<tr>
<td>4.1</td>
<td>Visual representation of three levels related to domain considerations</td>
<td>70</td>
</tr>
<tr>
<td>4.2</td>
<td>Summary of framework for domain analysis and corpus design</td>
<td>72</td>
</tr>
<tr>
<td>4.3</td>
<td>Domain analysis and corpus design for late nineteenth-century fiction novels by British and American writers</td>
<td>74</td>
</tr>
<tr>
<td>4.4</td>
<td>Domain analysis and corpus design for biographies</td>
<td>75</td>
</tr>
<tr>
<td>4.5</td>
<td>Domain analysis and corpus design for product manuals</td>
<td>76</td>
</tr>
<tr>
<td>4.6</td>
<td>Domain analysis and corpus design for external research grant proposals in the United States</td>
<td>77</td>
</tr>
<tr>
<td>4.7</td>
<td>Domain analysis and corpus design for White House press briefings</td>
<td>78</td>
</tr>
<tr>
<td>4.8</td>
<td>Domain analysis and corpus design for cooking shows</td>
<td>79</td>
</tr>
<tr>
<td>4.9</td>
<td>Domain analysis and corpus design for job interviews</td>
<td>80</td>
</tr>
<tr>
<td>4.10</td>
<td>Relationship among the target domain, the operational domain, and the corpus</td>
<td>99</td>
</tr>
<tr>
<td>4.11</td>
<td>Domain analysis and corpus design for the Academic Journal Research Corpus</td>
<td>107</td>
</tr>
<tr>
<td>4.12</td>
<td>Information about the PSYTB corpus from Miller (2012)</td>
<td>116</td>
</tr>
<tr>
<td>5.1</td>
<td>Number of texts for differing rates of prepositions (per 1,000 words) in the CORE corpus of online news</td>
<td>126</td>
</tr>
<tr>
<td>5.2</td>
<td>Number of texts for differing rates of prepositions (per 1,000 words), based on random samples from the CORE corpus of online news</td>
<td>127</td>
</tr>
<tr>
<td>5.3</td>
<td>Five random lexical growth curves for lemmas in the BNC</td>
<td>140</td>
</tr>
<tr>
<td>5.4</td>
<td>Lexical growth curve for lemmas in the online news sub-corpus of CORE</td>
<td>141</td>
</tr>
<tr>
<td>5.5</td>
<td>Frequencies of the 200 most frequent words in the BNC</td>
<td>143</td>
</tr>
</tbody>
</table>
List of Figures

5.6 Mean text range proportions across the top ten 1,000-word samples in the BNC, BNC Written, and BNC Written Biology 150
6.1 Graphical representation of our representativeness framework 157
6.2 Relationship among the target domain, the operational domain, and the corpus 159
6.3 Major corpus design factors that influence the accuracy of quantitative-linguistic analyses 160
6.4 Probability density distributions and means for entire domain (light grey; $M = 73.96$) and corpus (dark grey; $M = 64.56$) randomly sampled from operational domain #2 ($N = 1,000$), where $d = 0.33$ 165
6.5 Adjectives (per 1,000 words) in (1) full domain (mean plotted at horizontal grey line), (2) 100 random samples from experimental condition 1.1.1000 (light grey boxplot) and (3) experimental condition 2.1.1000 (dark grey boxplot) 166
7.1 Steps for creating a new corpus 179
7.2 Steps for evaluating the representativeness of an existing corpus 181
7.3 Summary of domain description for published academic writing 204
7.4 Summary of the operational domain for COCA Academic 206
7.5 Summary of the operational domain for BNC Academic 207
7.6 Summary of corpus design for COCA Academic 211
7.7 Summary of corpus design for BNC Academic 212
B1 Proportions of corpora according to their target domain, sampling unit, operational domain, and availability of a source list 264
B2 Proportions of corpora according to their stratification, proportionality, and randomness 266
### Tables

4.1 Domains included as case studies in this chapter  
4.2 Top-level discipline categories and example disciplines  
4.3 Taxonomy of academic research article types  
4.4 Disciplines and article types included in the operational domain for the AJRC  
4.5 Journals represented in the AJRC  
5.1 Mean scores and confidence intervals for a common linguistic feature and a rare feature  
5.2 Required sample size estimates for analyses of different linguistic features, based on a pilot corpus of online interviews  
5.3 Critical $RSE$ values for several error rates  
5.4 Descriptive statistics and $RSE$ for specific linguistic features in the CORE song lyric sub-corpus  
5.5 Average text proportions of the most common word types in the BNC, according to their frequency band  
5.6 Number of “target words” (out of a total of 10,000) correctly identified for each frequency band  
5.7 Comparison of the sets of lexical bundles identified in two random samples of texts matched for corpus size (extracted from CORE-News)  
5.8 Descriptive statistics for a subsample of L2 writing from BAWE for five linguistic features  
5.9 Descriptive statistics, CI, and required N for six categories of pronouns in online recipes  
5.10 Descriptive statistics and $RSE$ for grammatical complexity feature the L1 subsample of BAWE  
6.1 Designs of the twelve experimental corpora, manipulating the specification of the operational domain, sampling method, and sample size  
6.2 Average Cohen’s d values for the first three experimental conditions
List of Tables

6.3 Average Cohen’s d values for the second set of experimental conditions 168
6.4 Average Cohen’s d values for the third set of experimental conditions 169
6.5 Average Cohen’s d values for the fourth set of experimental conditions (averaged across 100 randomly extracted experimental corpora each compared to the full domain) 171
6.6 Average Cohen’s d values for the individual Wikipedia categories 173
7.1 Required sample size estimates for adjectives and adverbs, based on a pilot corpus of 100 Yelp restaurant reviews 189
7.2 Descriptive statistics and post hoc precision rates for adjectives and adverbs in the CRYRR, rank ordered from lowest RSE to highest RSE 191
7.3 Linguistic features from Biber (1988) Dimension 1 197
7.4 Required sample size estimates for twenty-six linguistic features in Biber (1988) Dimension 1, based on a pilot corpus of 100 YouTube vlogs 198
7.5 Descriptive statistics and post hoc precision rates twenty-six linguistic features from Biber (1988) Dimension 1, rank ordered from lowest RSE to highest RSE 200
7.6 Metadata for texts in BNC_AC and COCA_AC across subgenres, disciplines, and time 208
7.7 Descriptive statistics and post hoc precision rates for premodifying nouns and noun complement clauses in COCA Academic 213
7.8 Critical RSE values for several error rates 214
7.9 Descriptive statistics and post hoc precision rates for premodifying nouns and noun complement clauses in BNC Academic 214
Acknowledgments

The ideas and methods presented in this book have been in embryo for many years now. They were first conceived with Doug’s work on corpus design and representativeness in the 1990s, and new life was breathed into them during Jesse’s and Bethany’s graduate studies. The period of gestation from that time several years ago until now was longer than expected, but as we look back we are pleased to find that the extensive reading, research, piloting, drafting, and discussion that took place during those years was both necessary and fruitful. In one sense, the ideas in this book are very old, and we are indebted to the pioneers in statistical sampling who have addressed issues of sampling and representativeness either in theoretical terms or applied to other fields. In another sense, the ideas in this book are new – this is the first comprehensive framework for designing and evaluating language corpora for representativeness, and in order to accomplish that goal we have modified and adapted ideas from statistical sampling and other fields, and also developed novel methods, to fit the specific needs of corpus linguistics.

As we worked to develop new ideas and to successfully communicate old ones, we received helpful feedback and suggestions from students at Brigham Young University (fall 2014: undergraduate course on corpus linguistics; winter 2016: graduate course on corpus linguistics), Northern Arizona University (NAU) (fall 2016: undergraduate course on corpus linguistics; spring 2018: seminar on corpus linguistics and triangulation; fall 2018: undergraduate course on register variation; spring 2020: seminar on sampling in applied linguistics; “Coffee and Cronbach” statistics discussion group), and Iowa State University (fall 2020: seminar in corpus linguistics; spring 2021: corpus linguistics research group). Drafts and content of chapters from this book were used in all of these courses, and student feedback and discussion greatly improved the quality and clarity of this book’s content. The contributions of three now-graduated NAU students were invaluable: Daniel Keller was primarily responsible for the development and evaluation of the new corpora introduced in Chapter 7, Mike Suhan aided in the collection and analysis of the Wikipedia corpus, and Ally Kermad assisted with the collection of the information about corpora in Appendix B.
Acknowledgments

We are grateful to Helen Barton and Isabel Collins at Cambridge University Press for patiently encouraging us toward the finish line, with insightful guidance and feedback all along the way. We would also like to thank anonymous reviewers for Cambridge University Press who offered helpful comments on the proposal and an earlier draft of this book.