Jesse Egbert, Douglas Biber, and Bethany Gray, *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*, Cambridge: Cambridge University Press, 2022.

Ana Abigahil Flores Hernández
Universidad Autónoma del Estado de México
aafloresh@uaemex.mx

Pauline Moore
Universidad Autónoma del Estado de México
pmooreh@uaemex.mx

Corpus linguistics, the computer-aided analysis of very extensive collections of transcribed utterances or written texts (McEnery & Hardie, 2012), is an area of studies that has grown considerably during the past two or three decades. The core of this research methodology relies on the use of well-designed and selected language corpora, large collections of naturally occurring texts stored digitally (Egbert & Baker, 2020). Currently, corpus linguistics applications cover a wide range of studies about language such as discourse analysis, language acquisition, lexical and grammatical analysis, complexity, and evaluation of texts, among others. Hence, the process of design and compilation of a corpus needs to be conducted with great care to ensure reliable results. Representativeness in corpus design and evaluation is undoubtedly a key issue, despite the lack of agreement between scholars on a definitive concept that covers the full range of features involved in this notion.

The aim of this book is to introduce an operationalized concept of representativeness to be used in both the design and the selection of a corpus to meet your own research needs in the form of a guidebook with straightforward and easy-to-read explanations. The book is distributed into seven chapters taking the reader from an understanding of the basics of corpus design to the application of the guidelines. In the introductory chapter the authors establish an operational concept of corpus emphasising key characteristics, goals, and importance. Chapters Two and Three introduce the several approaches to representativeness and states a framework to conceptualise and determine its components. Chapters Four and Five describe thoroughly these two components and finally, Chapters Six and Seven conclude the discussion with a summary and some case studies to show how the proposed framework of representativeness can be applied in corpus design.

The First Chapter introduces several definitions of corpus, which are summarised to identify five main attributes, a corpus as a large collection of texts, a corpus as examples of natural or authentic language-in-use, a corpus as a representative sample of a particular language or domain, a corpus as an electronic and machine-readable resource, and finally, a corpus as a resource collected following specific principles that enables research about language. With all this in mind, the authors highlight the purpose of a corpus, "a generalizable empirical description of language use in a target domain" (p. 5), as the starting point to establish the importance of representativeness, highlighting the centrality of this notion in their final "operational" definition of a corpus as "a large, principled sample of texts designed to represent a target domain of language use" (p. 7). However, the authors stress that corpus linguistics should not be understood as the study of corpora, but rather as the study of patterns of language use in a language variety, as evidenced in the occurrence of such patterns in a corpus. They draw a useful parallel between the language variety under study as the population and the corpus as a sample of that population which serves as an underlying argument throughout the book. This necessitates the provision of sufficient detail on how representativeness has been addressed in the design stage of corpus building to meet the interests of corpus analysts and consumers of corpus research.

Representativeness has been variously defined in the field of statistics and they use the Second Chapter to present most of these approaches. Among other notions they look at the usage of the term representativeness as a positive evaluation of the data, the use of random selective procedures, a collection of typical or normal cases of language use, an attempt to cover the full range of linguistic variation, a miniature proportional sample of the population, a collection which provides a good estimation of parameters in the population as a whole from their frequency in the corpus, or that representativeness is linked to design for a specific research purpose. While most of these qualities would be considered intuitively attractive in corpus design, there are practical issues associated with achieving them. The authors provide guidance on practical random sampling, the typical uses of a word data should include, the coverage of the language major categories, the proportional distribution of the sample and the parameters found in them, in relation to the full target population and the relevance of research purposes. In addition, this chapter breaks away from the idea of size and balance as the main predictors of representativeness, closing with the assertion that this is a tailor-made estimation with reference to a particular research context.

In Chapter Three, the authors describe a methodological and conceptual framework for corpus representativeness and its operationalisation, outlining two main components of corpus design: the domain and the distribution. Domain considerations are related to the characteristics of the target real-world population and involve defining appropriate units and methods for sampling, and describing the operational domain of texts which constitutes the sampling frame. Distribution considerations deal with the size of the required sample and the internal linguistic variables which should be included to represent the target domain. These two components embody “the extent to which the corpus includes the full range of both text types and linguistic distributions in a domain” (p. 54) and reflect the central argument that the generalisation of sample estimations to the entire domain is the ultimate goal of corpus analysis. In this regard, however, they point out that representativeness is not an absolute quality, that is, corpora are not to be understood as either representative or not, rather they will have differing degrees of this quality and the design goal should be its maximisation.

Following on from this framework for representativeness, Chapter Four, “Domain Considerations", starts by pointing out that it is not possible to capture all the potential variation in language and that there are no catalogues to determine the definitive categories, strata, or full repertoire of existing text-types in the language. Accordingly, three steps should be followed to address representativeness in each domain: firstly, a general description of the domain, secondly, the specification of available texts and, thirdly, a procedure for sampling those texts. They describe each of these steps in detail, outlining connections between domain description and non-linguistic characteristics, in particular, external factors such as the population’s use of language and internal demographic or situational factors. The specification of texts allows for the identification of the main textual categories or strata, detecting the boundaries between categories, and available sources for those texts. The final step, sampling, entails deciding the proportion of texts to be included in each stratum, minimising bias from selection and coverage. The chapter also includes a description of the sampling procedures (i.e. stratification and proportionality) and closes with a case study about articles in academic journals to illustrate the methodology behind target domain composition.

In Chapter Five, “Distribution considerations", the authors elaborate on the question of how many texts should be included in a corpus to ensure representativeness. The objective is to facilitate the collection of representative samples by reliably capturing any quantitatively significant linguistic variable. To this end they discuss methods to determine the "corpus size required to measure linguistic rates of occurrence with high precision" (p. 124), offering a detailed explanation of linguistic variables and the statistical dangers of undersampling and oversampling. The chapter goes on to present statistical methods to determine precision and required corpus size for corpus building. They close with some practical cases taken from the British National Corpus (BNC) which highlight the fact that it is not possible (regardless of corpus size) to "guarantee occurrence of all words" (p.140).

Chapter Six puts the framework of representativeness presented in this book to the test by evaluating the accuracy of quantitative-linguistic analyses with reference to the corpus design factors of coverage bias, selection bias, and sample size (precision) using linguistic parameter estimation as the yardstick for representativeness. They achieve this by designing a set of 12 experimental corpora based on Wikipedia which manipulate the setting of the operational domain, use of sampling methods and corpus size. The results of their experiments allow them to conclude that both domain and distribution considerations work together for the accurate representation of quantitative-linguistic patterns.

The book concludes in Chapter Seven by summarising concepts and providing a list of key steps used to design new corpora and to evaluate existing corpora through the presentation of practical cases. This exercise is essential to show that their operationalisation of the term representativeness is not impractical as a requirement for corpus builders. To achieve this they look at the procedures to ensure representativeness in new corpora which aim to answer specific research questions about the use of language in restaurant reviews and vlogs and explain how to evaluate the suitability of large publicly available corpora, like the BNC and the Corpus of Contemporary American English (COCA) to describe the distribution of lexico-grammatical features of language use. This leads them to emphasize the importance of collecting and reporting clear data about the representativeness of the corpus with relation to a clearly identified operational domain.

This volume provides a practical definition of representativeness as well as the methodological framework to achieve it. It guides the reader-researcher through the process of selecting or designing a corpus for their research purposes and then offers statistical methods to test the decisions of the corpus compiler. The authors apply the theoretical framework in practical cases to show how those factors operate in real-world corpora. Useful features of the book are the simple and practical explanations and the organisation of the contents from less to more specialised concepts. Every chapter is introduced with a set of leading questions or “key issues" and closes with a summary of the most important points discussed or “key takeaways” as well as exercises in the application of the notions introduced for the design, evaluation, and use of corpora for research purposes. This book is a valuable guide for corpora users and designers, a must-read before beginning the process of corpora selection and design.

References

Egbert, Jesse, Paul Baker (2020) *Using Corpus Methods to Triangulate Linguistic Analysis*. London: Routledge.

McEnery, Tony, Andrew Hardie (2012) *Corpus Linguistics. Method, Theory and Practice*. Cambridge: Cambridge University Press.