ICAME Journal • 48(1) • 2024 • 109–112
DOI: 10.2478/icame-2024-0006

sciendo

ICAME Journal

Jesse Egbert, Douglas Biber and Bethany Gray. *Designing and evaluating language corpora: A practical framework for corpus representativeness*. Cambridge: Cambridge University Press, 2022. 300 pp. ISBN 978-1316605882

Review

Sharon Hartle

University of Verona, Italy

1 Background

Corpus linguistics, as a tool for research, is increasingly prevalent nowadays. It is relied on by researchers from diverse linguistics sub-disciplines, ranging from discourse analysis and sociolinguistics to computational linguistics, to name just a few. Designing and evaluating corpora that will be fit for the purpose they are being adopted for is imperative. This is particularly true since the need for robust and representative language corpora is well known. The authors of the book under review here, Jesse Egbert, Douglas Biber and Bethany Gray, have made significant contributions to the field of corpus linguistics, building on their research to advance the understanding and practice of corpus design and evaluation with a particular emphasis on representativeness, which dates back to Biber's work on corpus design in the 1990s (p. xii), honed over the years by means of their own extensive reading and research.

In their introduction (p. 3) to *Designing and evaluating language corpora: A practical framework for corpus representativeness*, Egbert, Biber and Gray underline the fact that in definitions of the term 'corpus', 'representative' is, indeed, one of the most frequent key concepts. Corpus representativeness refers to the extent to which a corpus reflects the full range of variability in the language or language variety that it is intended to represent. It is important for a number of reasons. First, it allows researchers to generalize their findings from the corpus to the larger population of language users. Second, it ensures that corpora are useful for a wide range of research questions, from those that focus on specific linguistic features to those that focus on more general aspects of language use. Two key aspects of representativeness that are explored in this volume are domain and linguistic representativeness, which the authors see as being central to the design and evaluation of representative corpora. This volume, therefore, addresses the need, in particular, to focus on representativeness and does so with both academic rigour and practical insight. Corpus representativeness, in fact, is crucial for ensuring the reliability and generalizability of research outcomes.

Achieving corpus representativeness, however, is not without its challenges. Several factors need to be considered, such as corpus size, composition, and balance. A larger corpus generally provides a more comprehensive and reliable representation of a language or language variety. However, the size of the corpus should be balanced with the available resources for data collection and analysis, together with the specific purpose for which that corpus is to be used. Careful sampling and consideration of

* *Corresponding author: Sharon Hartle, E-mail: sharon.hartle@univr.it*

Open Access. © 2024 Sharon Hartle, published by Sciendo.
This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

content focus, sources, and register are also key factors in the compilation of a corpus that adequately represents the target language and facilitates meaningful analyses across different linguistic phenomena.

Indeed, the book is written in an innovative, accessible style organised according to a systematic framework. In terms of layout, it is intuitively structured, with each section building logically upon the previous section. This progressive organization, coupled with clear subheadings and a wealth of examples, enhances the volume's value both as a clear introduction to the issues involved in compiling or assessing corpora as well as its potential implementation as a handbook. The inclusion of checklists, templates, and real-world scenarios transforms the text into a veritable toolkit for the corpus designer. The graphical representations and tables throughout the volume offer visual summaries of complex concepts, aiding in the reader's comprehension and retention of the material.

Each chapter begins with an introduction of the main topics that will be encountered in the chapter and concludes with summaries of 'Key Takeaways' from the discussion and a series of exercises. This design makes the volume ideal for use both with postgraduate students on Higher Educational (HE) courses and with researchers who are new to the field. The readers are led through a series of considerations, and by means of both the reflective and practical exercises they will be able to set about designing and evaluating their own representative corpora. The target audience, however, also includes those who need to read corpus linguistics research findings with a critical eye.

The volume's central thesis emphasizes the paramount importance of representativeness in corpus design – a factor that critically influences the reliability of linguistic analyses and the validity of conclusions drawn from corpus data. Egbert, Biber and Gray meticulously dissect representativeness, offering a nuanced perspective that moves beyond the mere size or diversity of text types. They advocate a stratified approach to corpus compilation that accounts for variation across genres, modes, and demographic factors, a standpoint that is increasingly resonant in current linguistic research.

2 Criticisms of corpus linguistics

The question of where to position the corpus linguistics approach in research, within the study of language and within scientific approaches to the study of language, is in itself problematic (McEnery and Brezina 2022). Teubert (2005) and Stubbs (2006) argue that corpus linguistics is a theoretical approach, whereas Gries (2009) and McEnery et al. (2006) see it as a concrete methodology applied to the investigation of linguistic research questions. In order to combat such limitations, Egbert, Biber and Gray propose a practical framework for corpus representativeness that provides researchers with step-by-step guidance and examples for designing representative corpora, in line with their inquiries.

One of the fundamental criticisms presented concerns the representativeness of corpora. Critics argue that no corpus, no matter how vast, can adequately capture the full spectrum of a language (Widdowson 2000). This limitation becomes particularly significant when considering dynamic aspects of language, such as slang or evolving jargon, which may not be sufficiently represented in existing corpora (Biber and Reppen 2002). Another criticism which is levelled in the field concerns the lack of consensus regarding corpus representativeness and sample collection methodologies. Additionally, corpus-based research may be subject to biases and limitations inherent in the data itself, such as the overrepresentation of certain genres or registers.

3 Structure of the book

The early chapters in the book lay a foundational understanding of key principles in corpus linguistics, setting the stage for the subsequent methodical exploration of corpus design. Chapter One provides an introduction to the term 'corpus', discussing in particular why the design, evaluation, and representativeness are key concepts for corpus compilers, analysts, and corpus consumers. Chapter Two expands on the concept of corpus representativeness, documenting conceptualizations of the term both from the

viewpoint of statistical/sampling theory and in the history of corpus linguistics itself. Questions are explored, such as the extent to which there is a shared understanding of representativeness in statistics, social sciences, or specifically in corpus linguistics, or the question of how corpus size is related to representativeness. Chapter Three goes on to introduce the framework of 'corpus representativeness' presented in the following chapters, taking into consideration both conceptual and methodological aspects. This chapter investigates the relationship between linguistic research goals, corpus design, and corpus representativeness and outlines the two main components of corpus representativeness: domain and distribution considerations.

It is in the next section, which explores further the methodology, however, that the book perhaps truly comes into its own. The authors present a step-by-step guide that is both theoretically informed and practically oriented, providing readers with the tools necessary to craft corpora that are fit for their intended linguistic inquiries. Chapter Four looks at domain considerations, or rather the degree to which a corpus includes the full range of text types in a domain. This chapter includes the process of describing the target domain of language use as the basis for corpus design and evaluation. Chapter Five complements this as it turns to considerations of distribution and methods of evaluation regarding corpus size. The chapter addresses the important question of how large a corpus needs to be if it is to provide accurate estimates of the extent to which a corpus reliably captures the range of distributions for different linguistic features. Consequently, these two chapters can be thought of as investigating in greater depth the two main pillars of representativeness introduced earlier in Chapter Two.

The concluding chapters bridge the gap between theory and practice by walking the reader through a series of case studies that explore the ways in which domain and distribution considerations work together to form the basis of corpus representativeness. These case studies serve as examples of the book's practical framework in action, offering a glimpse into the complexities of real-world corpus construction and evaluation.

Chapter Six revisits the various components of the framework and shows how these may combine into a single, unified approach to achieving a high degree of corpus representativeness. This is explored with the example of an experimental design for a corpus consisting of Wikipedia articles, illustrating some of the ways in which domain and distribution impact the accuracy of linguistic findings, depending on the corpus design.

Chapter Seven, written in collaboration with Daniel Keller, deals with the application of the corpus representativeness framework in two specific contexts: a researcher designing their own corpus and a researcher evaluating an existing corpus in order to use it for their own goals.

In conclusion, an important aspect of the book is that it does not shy away from the challenges and trade-offs inherent in corpus design. The discussion on the balance between comprehensiveness and feasibility is particularly enlightening, offering a candid look at the constraints of time, resources, and access to language data that researchers commonly face.

Designing and evaluating language corpora is commendable for its clarity, depth, and practicality. It stands as a testament to the authors' expertise and their dedication to advancing the field of corpus linguistics. It can be considered to make a substantial contribution to the corpus linguistics literature. It is a volume that not only underscores the necessity of representativeness in corpus design but also empowers readers to achieve it. The book's clear layout and accessible content make it an exemplary guide for those embarking on the complex yet rewarding journey of creating balanced and meaningful linguistic corpora.

References

Biber, Douglas and Randi Reppen. 2002. What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition* 24 (2): 199–208.

Gries, Stefan. 2009. What is corpus linguistics? *Language and Linguistics Compass* 3 (5): 1225–1241.

McEnery, Tony and Vaclav Brezina. 2022. *Fundamental principles of corpus linguistics*. Cambridge: Cambridge University Press.

McEnery, Tony, Richard Xiao and Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. London: Routledge.

Stubbs, Michael. 2006. Corpus analysis: The state of the art and three types of unanswered questions. In G. Thompson and S. Hunston (eds.). *System and corpus: Exploring connections*, 15–36. Sheffield, U.K.: Equinox.

Teubert, Wolfgang. 2005. My version of corpus linguistics. *International Journal of Corpus Linguistics* 10 (1): 1–13.

Widdowson, Henry George. 2000. On the limitations of linguistics applied. *Applied Linguistics* 21 (1): 3–25.