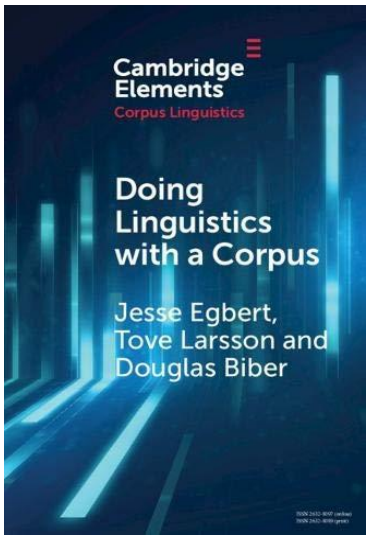
Jesse Egbert, Tove Larsson, Douglas Biber

Doing Linguistics with a Corpus

Methodological Considerations for the Everyday User



Cambridge, Cambridge UP, 2020, pp. 80

Review by Elena Mattei

Keywords: *corpus linguistics, research process, statistical measures, critical qualitative analysis, reliability*

Research in linguistics studies has increasingly relied on computational and corpus methods in the last decades, in order to provide both reliability and generalizability to qualitative linguistic results. This need for quantitative findings and statistical procedures has been accompanied, especially in the last ten years, by a growing awareness of the complexity and constant expansion of computational tools. The consequence of this is a call for researchers in every field to be guided on ways in which they can conduct corpus analysis and provide systematic, replicable interpretations of patterns of language use.

This book falls within the literature on methodological approaches, which helps scholars engage practically with corpus research in linguistics. Building on previous similar works (Biber et al. 1998; McEnery et al. 2006), the authors provide a step-by-step guide into corpus research design and processes, implementing a problem-oriented approach that fosters the acquisition of practical skills to deal with the vast range of corpus tools, objects and statistics available today. This volume in particular aims to provide scholars engaging with corpus analysis with some understanding and, in addition, training on the processes that lie behind statistical techniques, technological resources and corpus tools/methods. As the authors state in the introductory chapter, "some understanding of how a car works can be a useful complement to the simple practice of getting behind the wheel and turning the key" (1). To achieve this aim, the authors

structure the book in eight chapters, which, apart from the introduction (Chapter One) and the conclusion (Chapter Eight), introduce the chronological steps in the research process for a discourse analyst. Chapter Two and Three explain how to choose a suitable corpus and research design according to the research question and object of study, whereas Chapters Four to Six present a series of methodological issues related to the right choice of both statistical and computational tools, raising awareness of the existence of bias and limitations behind these procedures. Chapter Seven acts as the conclusion of the research process and emphasizes the importance of providing meaningful qualitative insights into quantitative findings. Finally, Chapter Eight summarizes the suggestions provided throughout the book. This volume thus explores the ways in which a critical review of the tools, research objects (such as a language variety or register) and possible statistical measures can be conducted practically, reinforcing the notion that each step of a study, no matter how simple it may seem, deserves critical attention and informed, reasoned choices. It therefore makes a valuable contribution to the field which will be useful both for researchers at the beginning of their careers and experts alike.

Chapter Two is called "Getting to Know Your Corpus" and represents the first methodological issue to tackle when undertaking corpus analysis. This section is fundamental as it explains how to build an adequate dataset according to the research question posed, among a huge array of options when it comes to language corpora (4). The chapter presents then a case study that analyzes the generic composition and distribution of specific linguistic variables across different corpora, in order to evaluate the degree of representativeness of the target domain language and the generic balance of the corpus. This practical example enables the readers to understand the potential limitations of data and the wrongful interpretations that can derive from particular choices.

Chapter Three focuses on the definition of *research design*, or the principled collection and organization of the data according to the research question in mind. It is a chapter which might be of particular interest to novice researchers who—according to the authors—tend to collect and investigate texts without a clear research question, thereby ending up drawing the wrong conclusions on the variation and occurrences of specific linguistic features. This section in particular aims to make the readers aware of the difference between variationist and descriptive studies, which have different observable units, variables and objectives and therefore outcomes. As a matter of fact, while variationist linguistics attempts to discover how a linguistic entity varies according to its adjacent items, thereby investigating how a particular context affects the choice of grammatical constructions, descriptive approaches focus on the rate of occurrence of different linguistic attributes to detect a register or a discourse (19-20).

Chapter Four introduces readers to the nature and the limitations of each corpus tool in detail, in order for him or her to be able to select the most suitable one for his or her specific goals, and to check each result by qualitative analysis, to further corroborate the numbers. A complete reliance on computed measures, in fact, can lead to misleading results and consequently misinterpretations. Most statistical tools employed to calculate the keyness of a collection of texts, in fact, might report findings which are not representative of a specific researched discourse (30-32).

Chapter Five continues the line of reasoning presented in Chapter Four and focuses on a critical approach to the limitations of fast-processing computational tools, by describing the potential inaccuracy and irrelevance of pure statistical findings. This reinforces the need of researchers for further qualitative analysis.

Chapter Six concludes the discussion on the excessive reliance on computational tools by prompting the readers to choose the most appropriate statistical method for their own research and to review each statistical result thoroughly "to draw reliable and meaningful conclusions about language use" (40). The authors warn against over-reliance on *null hypothesis significance testing* in the assessment of differences between the mean rates of occurrences of particular features in large corpora, showing how this measure, because it is sensitive to sample sizes, will always find statistical differences which appear to be significant, without taking real effect size into consideration.

Finally, Chapter Seven explores ways of providing qualitative insights into language use and discovering meaningful patterns by drawing upon three main sources of information: the linguistic context (with the concordance tool), the linguistic theories and the text-external context, which includes both precious metadata such as the demographics of a target population, the registers and distribution of the sources and the previous findings.

In conclusion, this book is an essential tool for any scholar working with quantitative data in corpus linguistics, due to its clear approach and structure of chapters that follow the chronological steps involved in corpus analysis research design. The practical focus on solving often overlooked methodological issues with simple—but not simplistic—language make this guide useful and readily applicable by both inexperienced researchers, who might find themselves overwhelmed by the vast range of data and statistical/computational tools available, and accomplished scholars who want to conduct their research from a fresh perspective. The type of knowledge and especially awareness offered by this book will help researchers avoid potential bias, and inappropriate tools or measures whilst reinforcing the need for qualitative interpretation as well as quantitative analysis.

This volume therefore ensures that the readers have the necessary instruments to critically work on each step of the corpus analysis without having to master each statistical or annotation tool or developing their own computational knowledge. As also McEnery and Hardie argued, linguists who use corpus data are not expected "to become fully competent in computer programming […] or in the more complex statistical analyses" (2012, 226). They should work instead towards a conscious implementation of these procedures and statistical measures, which is what this volume advocates. However, due to the focus on the research process, this book does not provide an exhaustive description of the tools available, nor a section on the existing theoretical approaches to corpus analysis. Therefore, the reading of this book is highly recommended alongside a more classical introduction to Corpus Linguistics.

Elena Mattei *is a PhD candidate in Digital Humanities and English Language at the University of Verona. Her research interests focus on the collection and analysis of tourism multimodal corpora on Social Media, with a particular focus on both Visual Design and Systemic Functional Linguistics theories.*

Works cited

Biber, Douglas, Susan Conrad and Randi Reppen. *Corpus linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.

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

McEnery, Tony, Richard Xiao and Yukio Tono. *Corpus-based Language Studies: An Advanced Resource Book*. Oxfordshire: Taylor & Francis, 2006.