Data and Methods in Corpus Linguistics: Comparative Approaches

Ole Schützler & Julia Schlüter (Eds.)
Cambridge: Cambridge University Press, 2022. 358 pages.
ISBN: 978-1-108-49964-4

Corpus linguistics plays a critical role in contemporary scientific research. It has a wide range of applications, and the advancement of digital technologies means that our capacity to store and process large volumes of text has increased to an extraordinary degree. Corpus linguistics is now widely used in language research, translation studies, and other disciplines. Although many relevant research methods have emerged, many questions remain concerning methodology. It can be said that *Data and Methods in Corpus Linguistics* is a timely work to provide guidance for researchers. Starting from comparisons, it discusses the problems that arise from the perspective of dual methods/data sets so that the advantages and disadvantages become apparent, which should help researchers make better methodological choices.

This book contains four main chapters, which provide guidance concerning methods of analysis for all scholars interested in corpus linguistics. It incorporates discussion of case studies, uses comparative analysis, analyzes the strengths and weaknesses systematically and objectively, uses concise language, draws conclusions, and makes recommendations. In addition, it also uses more advanced technologies to focus on possible future data sets, machine learning, and other issues. As an emerging interdisciplinary research branch, corpus linguistics can not only be used as a tool to provide objective and quantitative data (Atar & Erdem, 2019), but is now finding new applications in the era of AI and machine learning.

Following the case study approach, the first section in Chapter 1 mainly revolves around two types of databases: Google's large-scale database GBN and two standard corpora, the British National Database (BNC) and the US National Database database (COCA) for comparative analysis. Lukas Sönning and Julia Schlüter conducted experiments focusing on the size of the database and the source of metadata, then summarized the advantages and disadvantages of the two databases: GBN has an extensive range of

 Ibérica 47 (2024): 332-335
ISSN: 1139-7241 / e-ISSN: 2340-2784
https://doi.org/10.17398/2340-2784.47.332

data, it can collect more tokens for each lexeme (p. 39) and can make the results of sample analysis more stable. Second, its broad data types make it more suitable for studying the differences between languages, which cannot be done with small data sets. The different information obtained by using BNC and COCA is highlighted. The authors also point out that combining these two databases can help people understand research data more comprehensively (p. 4).

The second section is centered on a specific case: the "Principle of Rhythmic Alternation" (PRA). This section points out that PRA is often considered a driver of different types of phenomena. Its properties require different corpus approaches, and different types of data are involved. The impact of data and method interactions are presented at the end of the section. For example, under the premise of using corpus data, written or orthographically transcribed spoken data will be more widely used than spoken data with access to audio recordings (p. 65). The research results also show that the 'idiom principle' or the concept of 'lexicalized sentence stems' may be relevant for phonological levels (p. 67). For this reason, it is a good choice to combine different data sources and methods for research, which echoes the conclusions from the section on using both large and small databases.

In Chapter 2, the author puts forward two questions: 'what goes into a corpus', as well as 'what goes into an analysis'; the corpus is a huge collection of texts, and it is also a sample (Biber, 2011). However, the corpus is not equal to the language, no matter how many sentences it contains (Jones & Waller, 2015). So, the hierarchical structure of the corpus needs to be taken into account (p. 4). The main issue discussed in this chapter is the preparation of corpus data in research and how to analyze it. This chapter is divided into 3 parts. The first is Fabian Vetter's part, which explains the design of experiments to explore differences between corpus registers. The analysis also proves that differences between them may be due to the different sampling strategies adopted by the corpus compiler. In the compilation of future corpora, Fabian Vetter recommends adding situational characteristics to annotated texts (p. 98). The second section is about the passive voice. Alternately selecting different baselines, Sean Wallis and Seth Mehl stress that normalized frequencies often fail to yield meaningful measures (p. 5). Therefore, a baseline indicating opportunities of use is vital to make the data reliable. At the end of the section, the advantages and disadvantages of three different baselines are

provided. The last section is by Lukas Sönning and Manfred Krug, who call for richer metadata and elaborate on the benefits of linking corpus data to speakers.

The main content of Chapter 3 concerns how various researchers utilize statistical methods to evaluate the influence of specific factors on context, and assesses their advantages, disadvantages, and limitations. In addition, the article also develops regression analysis and distance-based visualization. For example, in the first section, Tobias Bernaisch compares, among others, the Generalised Linear Mixed-Effects Models, and concludes that even for the same set of data, if different models are used, there will be differences and diverse observation results. For another example, in the third section, Natalia Levshina compares the standard frequencies with the recent Bayesian method. She focuses on multiple logistic regression with mixed effects and believes that Bayesian statistics could effectively stabilize data and reduce the amount of data preparation. This could provide a powerful tool to overcome limitations and promote collaboration (p. 8).

The last chapter concerns the combination of corpus linguistics and computer and machine learning. Because the content of the corpus can be processed using different types of annotations, the specific combination of computer and machine learning will produce innovative methods to serve better the tasks required by corpus linguistics. For example, in the first section, Gerold Schneider uses a corpus-driven approach to study English grammar changes, identifying words combined in a specific form or trend in grammar (Biber et al., 2010). It is clear that corpus linguistics is not an independent discipline. It can draw inspiration from other fields to generate new research hypotheses.

Overall, this book is a useful contribution to the development of corpus linguistics. Through different research cases, this book adopts different types of experimental methods and discusses the basic principles behind these methodologies in depth, so that everyone who wants to participate in corpus linguistics research can distinguish the pros and cons of different corpora, or their data processing methods, and make better choices. Readers can gain an in-depth understanding of essential issues and the latest research methods in related fields through this book. One positive aspect is that the disciplines of corpus linguistics and computer science are combined in the second half of the book, providing researchers in linguistics and other disciplines with rich research ideas and methods. This book is an indispensable reference

book in the field of corpus linguistics and has significant value for scholars, teachers, and students engaged in language research.

Received 23 April 2023
Accepted 24 September 2023

Reviewed by **Ying Zou**
Guangdong University of Technology (China)
z1210155271@gmail.com

References

Atar, C, & Erdem, C. (2019). The advantages and disadvantages of corpus linguistics and conversation analysis in second language studies. *IX Scientific and Practical Internet Conference of Young Scientists and Students* (pp. 140-146). Young Scientists Council of the National Academy of Educational Sciences of Ukraine.

Biber, D. (2011). Corpus linguistics and the study of literature: Back to the future? *Scientific Study of Literature*, *1*(1), 15-23. https://doi.org/10.1075/ssol.1.1.02bib

Biber, D., Reppen, R., & Friginal, E. (2010). Research in corpus linguistics. In R. B. Kaplan (ed.), *The Oxford handbook of applied linguistics*. Oxford Academic. https://doi.org/10.1093/oxfordhb/9780195384253.013.0038

Jones, C., & Waller, D. (2015). *Corpus linguistics for grammar: A guide for research*. Routledge.