RiCL Research in Corpus Linguistics

Review of Barth, Danielle and Stefan Schnell. 2022. *Understanding Corpus Linguistics*. London: Routledge. ISBN: 978-0-429-26903-5. DOI: https://doi.org/10.4324/978042926903

Isabel Zimmer – Elen Le Foll
University of Cologne / Germany

Understanding Corpus Linguistics provides an introduction to corpus linguistics and its application in multiple areas of linguistics. Written with undergraduate and graduate students of linguistics in mind, the textbook outlines how corpora can improve our knowledge of languages by providing authentic data. The book is divided into three parts: the first part provides an overview of basic concepts of corpus linguistics, the second part focuses on the processes of working with and creating corpora, while the third part explores the contribution of corpora to a selection of sub-disciplines of linguistics.

The opening chapter outlines the focus of corpus linguistics by differentiating it from other methods used in linguistics, such as acceptability judgments and experimental setups. It weighs up the advantages and disadvantages of corpus-linguistic methods. Barth and Schnell further illustrate the diverse applications of corpus linguistics in various sub-disciplines of linguistics, including sociolinguistics, psycho- and neurolinguistics.

Chapter 2 provides a comprehensive overview of the basic terminology of corpus linguistics, for instance, elaborating on the difference between word forms and lexemes, and types and tokens. It briefly explains how situational and/or text-internal contexts can influence the use of linguistic forms. Furthermore, the authors address the fact that corpora can only ever reflect a subset of language use, sampled from a relatively limited portion of the population.

The third chapter focuses on corpus composition and corpus types. Barth and Schnell emphasise that representativeness and a balanced number of text types are essential for corpora as an empirical data basis, and further highlight the importance of metadata, hereby highlighting potential issues with web corpora.

The objective of Chapter 4 is to demonstrate that corpus linguistics can be applied to many subdisciplines, for instance for questions in morphology or phonetics, discourse or sign language. For each subdiscipline, one or two studies are presented.

Although Chapter 5 is entitled “Corpus Queries”, it does not introduce queries as such, but rather different methods for analysing corpus data. This is likely explained by the fact that the authors refrain from explaining the use of any specific tool or programming package to avoid the textbook quickly becoming outdated. That said, the chapter does list some example corpus tools, *R* packages, and *Python* modules in standalone textboxes, which help to concretise the explanations. There is a strong focus on quantitative corpus linguistic methods with surprisingly little emphasis on concordancing, which is mentioned only briefly after frequency lists, keywords, dispersion plots, Zipfian distributions, collocations and bigrams, and association measures. The chapter ends with an introduction to regular expressions for corpus querying.

Chapter 6 begins by introducing three corpus building scenarios that allow the authors to cover a range of typical situations in the space of a short chapter. The chapter addresses the identification, selection, and evaluation of texts for corpus compilation, the collection procedure, copyright and privacy, as well as technical considerations such as how to deal with different modes and scripts. It provides more detailed information about the transcription of spoken data, including how to link transcripts to raw data using *ELAN* (2020) and concludes with short sections on data formats, the inclusion of metadata, and how to publish a corpus.

In Chapter 7, which deals with corpus annotation, Barth and Schnell provide brief descriptions of different types of annotation, from phonetic and prosodic annotation to discourse and reference annotation, including morphological and semantic annotation and part-of-speech (POS) tagging. The second half of the chapter focuses on corpus annotation in the context of cross-linguistic typological research with concrete examples from *The Social Cognition Parallax Interview Corpus* (SCOPIC; Barth and Evans

2017) and the *Multilingual Corpus of Annotated Spoken Texts* (Multi-CAST; Haig and Schnell 2015).

Chapter 8 begins by introducing foundational concepts of statistics for corpus linguistics including sampling, dependent and independent variables, distributions, range, and spread. It continues with worked examples of the use of chi-square and correlation tests, mixed-effects logistic regression, classification tree, and random forests applied to real-life corpus data. It also includes shorter sections on clustering and on how to report the results of statistical tests.

Following a brief introduction to sociolinguistics, Chapter 9 explains how corpora are used in the study of dialect and regional variation and dialectometry. It outlines the types of variables typically included in sociolinguistic studies and how corpus analyses can inform our understanding of variation and language change.

Chapter 10 focuses on language documentation and its use of corpora. Despite the smaller size and limitations of many corpora in this field, the authors emphasise their crucial role in preventing language loss. They elaborate on the process of corpus building in language documentation and provide examples of annotations of different data types. Due to the limited size of the corpus, research questions must be adapted, and only specific objectives can be addressed, in contrast to those that can be analysed with larger corpora. Despite the limited corpus size, the authors exemplify different analyses that can be conducted and conclude this chapter by discussing the limitations and advantages of smaller corpora.

Chapter 11 introduces corpus-based typology. The authors discuss some of the assumed language universals, demonstrating that there is, in fact, considerable variation across different languages. Linguistic diversity is exemplified with the use of different expressions for referential entities. To conclude, the authors outline issues and biases in corpus-based typology research, particularly when working with corpora that consist mainly of written data.

Understanding Corpus Linguistics is written in accessible language. Even relatively basic terminology is explained in detail so that beginners —the target readership of the textbook— are well catered for in this respect. The authors cover a lot of material in a relatively short textbook. Each chapter begins with a list of keywords

and concludes with some recommended further readings. The exercises, which build on what has been previously explained, are also a great addition to the textbook.

Whilst we recognise that there is no perfect order that will suit all readers, we found that the order of some of the chapters was not the most intuitive for us. In particular, we believe that the order of Chapters 5–7 (corpus queries → corpus building → corpus annotation) may be difficult for corpus novices. Whichever order is chosen (and the chosen order certainly has its justifications), cross-references between the chapters would help the reader to find the information they need more easily. A real strength of the textbook is that it mentions many different corpora, representing a wide range of languages and designed for use in different subdisciplines of linguistics.

As the title suggests, this textbook is about *understanding* corpus linguistics, not necessarily about *doing* corpus linguistics. As a result, Chapter 7, for example, focuses explicitly on the types of annotation and annotation schemes used in corpus linguistics, independently of any software or tool. With this in mind, this focus on tagsets rather than POS taggers makes sense, but it does mean that the reader is left with no idea as to how to actually perform a task as simple as POS-tagging, which may prove frustrating for some.

Chapter 8, on statistics for corpus linguistics, contains some inaccuracies. On p. 138, it states that "parametric tests need to meet assumptions like following a normal distribution and independence. Otherwise, non-parametric tests need to be used." This statement suggests that non-parametric tests can be used when the assumption of the independence of the data points is violated. This is not the case for most non-parametric tests used in corpus linguistics, for which the independence assumption still holds (for instance, the Mann–Whitney U test or the Kruskal-Wallis test). Most problematically, on p. 152, we read that "the *p*-value represents the chance that the null hypothesis would be true if we observed this sample of data." This definition of *p*-values is incorrect. Although they are often misconstrued as such, *p*-values do not correspond to the probability that the null hypothesis is true or false (see, for instance, Winter 2020: 171). Rather, *p*-values correspond to the probability of observing a result as extreme as, or more extreme than, the one obtained from the sample, if the null hypothesis were true. The "Further Reading" section of this chapter lists several books that provide excellent introductions to statistics for linguistics (including Winter 2020) and that are at an appropriate level for the target readership of the textbook. In contrast, we fear that the

advice to "just start reading [online] forums and seeing what you can glean [...]" (p. 163) is less sensible and unlikely to lead to sound statistical literacy among student readers.

Throughout the textbook, the authors place a commendable emphasis on the reproducibility and replicability of corpus linguistics research. For instance, in Chapter 5, they stress the importance of documenting workflows, which is rarely mentioned in corpus linguistics textbooks. Barth and Schnell also innovate by mentioning two studies which some would consider to be 'null results' (Chapter 9), but which we agree are very much still worth reporting about. It is also refreshing to see that the authors include the publication of a corpus as a:

> definitional feature of any corpus [...] because the primary purpose of a corpus is serving as a source for linguistic research, and being data (literally meaning 'given') entails that language scientists should be given the opportunity to look at the same things (the 'data'), including the surrounding context, when evaluating linguistic analyses and respective theories (p. 93).

While this is a very welcome addition to the definition of a research corpus, the link between the section on the "Publication of the Corpus" (6.6) and on the "Availability of Texts: Copyright and Privacy" (6.6.2) could be made clearer. It feels somewhat of an understatement to claim, on page 97, that "even some texts available through the internet may have some copyright protection or restrictions on usage." Given that the idea of publishing corpora has not yet been fully embraced by the corpus linguistics community, it might be worth including some examples of repositories where corpora can be made available. The authors choose the example of the Multi-CAST corpus, which is worth mentioning for several reasons, but which is published on a dedicated corpus website. We have our doubts as to whether this is the best example of how to share a corpus. Corpus websites need to be maintained and, as many older projects have sadly shown, links quickly become broken, resulting in the loss of valuable corpus resources. In addition, building an entire website may seem overwhelming to many researchers. For both these reasons, we suggest mentioning open repositories —for instance, CLARIN,[1] OSF,[2] TrolLing[3] or Zenodo[4]— which provide more sustainable infrastructures for corpus sharing with quick and easy uploading procedures.

[1] https://www.clarin.eu/content/data
[2] https://osf.io/
[3] https://dataverse.no/dataverse/trolling
[4] https://zenodo.org/

In conclusion, *Understanding Corpus Linguistics* covers a lot of ground, while making complex concepts of corpus linguistics genuinely digestible for undergraduate and graduate students. Although we picked up on a few problematic passages, particularly in the statistics chapter, they do not overshadow the book's overall value. What sets this textbook apart is its commendable emphasis on providing examples in languages other than English, including signed languages, a very welcome addition to the existing corpus linguistics literature, which has so far had a very strong focus on English as the object of study. The textbook is also pioneering in its strong focus on typology, while at the same time offering interesting insights into how many other subdisciplines of linguistics can benefit from corpus research. By addressing these neglected areas, this textbook effectively fills a conspicuous gap in existing corpus linguistics textbooks, making it a valuable resource for linguistics students and educators alike.

References

Barth, Danielle and Nicholas Evans. 2017. SCOPIC design and overview. In Danielle Barth and Nicholas Evans eds. *LD&C Special Publication No. 12: The Social Cognition Parallax Interview Corpus (SCOPIC): A Cross-linguistic Resource*. Honolulu: University of Hawai'i Press, 1–23.

ELAN. 2020. (Version 6.0) [Computer software]. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. https://archive.mpi.nl/tla/elan.

Haig, Geoffrey and Stefan Schnell eds. 2015. *MultiCAST (Multilingual Corpus of Annotated Spoken Texts)*. https://multicast.aspra.uni-bamberg.de. (23 December 2024.)

Winter, Bodo. 2020. *Statistics for Linguists: An Introduction Using R*. London: Routledge.

Reviewed by

Isabel Zimmer
University of Cologne
CRC 1252 "Prominence in Language"
Luxemburger Str. 299
50939 Cologne
Germany
E-mail: ifuhrma2@uni-koeln.de

Elen Le Foll
University of Cologne
Department of Romance Studies
Universitätsstraße 22
50937 Cologne
Germany
E-mail: elefoll@uni-koeln.de