# Understanding Corpus Linguistics by Danielle Barth & Stefan Schnell, 2022

Zahra Ghane & Mehrdad Vasheghani Farahani



Suggested Citation:
Ghane, Z., & Farahani, M. V. (2023). Understanding Corpus Linguistics by Danielle Barth & Stefan Schnell, 2022: Routledge 2022, New York, ISBN 9780367219628 [Review of Understanding Corpus Linguistics by Danielle Barth & Stefan Schnell, 2022: Routledge 2022, New York, ISBN 9780367219628]. Corpus Pragmatics, 7(3), 291–295. Springer. https://doi.org/10.1007/s41701-023-00145-y

Corpus Pragmatics (2023) 7:291–295
https://doi.org/10.1007/s41701-023-00145-y

BOOK REVIEW

# Understanding Corpus Linguistics by Danielle Barth & Stefan Schnell, 2022

## Routledge 2022, New York, ISBN 9780367219628

**Zahra Ghane**[1] · **Mehrdad Vasheghani Farahani**[2]

Received: 25 February 2023 / Accepted: 28 April 2023 / Published online: 19 May 2023


Contemporaneously with the advances of technology as well as the advent of computers in language studies, we have witnessed a boom in the emergence of new books in Corpus Linguistics (see for example Dash & Ramamoorthy 2019; Paquot & Gries, 2020; Seoane & Biber, 2021). From among the informative books in this fast growing field of knowledge is the current one authored by Barth and Schnell in 2022. This work of scholarship has been organized in 11 chapters, which provide readers with state-of-the-art concepts of theory and practice for conducting research in the domain of Corpus Linguistics.

The first two chapters function as an introduction in which the authors, succinctly, shed some light on the basic concept of corpus, its divergence from other approaches as well as its convergence with other usage-oriented fields within linguistics such as Sociolinguistics, Linguistics Typology and Language Change. The authors provide the reader with a definition of corpus and Corpus Linguistics, words, lexeme, type and token as well as some basic statistical concepts such as mode, mean and median. Later on, the authors make a distinction between structural context, syntagmatic context and constructional context in order to delineate the role of context in corpora.

There are different types of corpora with specific composition criteria, which need to be delineated for the readers. In this regard, Chapter three, which is thematically divided into two parts, is a detailed description of the corpus composition criteria and typology. In the first part, the authors enumerate such concepts as size, balance, representativeness as well as authenticity and spontaneity as the core criteria for compiling a corpus. Furthermore, a subtle distinction is made between raw, primary

✉ Zahra Ghane
linglistsubscribe@gmail.com

[1] Department of American and English Studies, Heinrich Heine University, Düsseldorf, Germany

[2] Department of Applied Linguistics and Translation Studies, Leipzig University, Leipzig, Germany

and metadata. Primary data that subsumes metadata is "recording of speech event or a written text, including paralinguistic properties (…) metadata are the searchable text data of a corpus in a written form in digital format" (pp. 32–33). In the second part, different types of corpora such as national, monolingual, multilingual, reference, monitor, learner, spoken, written, parallel, web, diachronic, synchronic, and annotated corpora, including tagged corpora and treebanks with vivid exemplars, are introduced.

In order to be able to conduct solid corpus-based research, it is essential to have a profound understanding of corpus query tools and techniques (Dash & Arulmozi, 2018). For this reason, the authors equip the readers with a wide range of necessary corpus queries including type-token ratio, frequency plots, key words, dispersion plots, ZIPFIAN distribution, concordance lines and keywords in context and N-grams in Chapter five. Each of these concepts is clarified with vivid examples and detailed information in such a way that by reading the chapter thoroughly, the reader will come to grips with corpus queries, how to engage with and run them and how to make tangible conclusions. At the end of this chapter, the authors make the first foray into regular expressions and specialized query languages.

With the tremendous advances of computer science, we have witnessed an upsurge of already available corpora; however, it is sometimes inevitable to create do-it-yourself-corpora. Regardless of the corpus type, the researcher has to know the nontrivial principles of creating corpora to a considerable degree. In this regard, Chap. 6 focuses on practicalities of corpus building and the wide range of various steps, which must be taken into consideration including selection and collection of texts, designing the corpus for specific goals, adding metadata, copyright and privacy, transcribing the spoken corpus and data format for corpus creation.

Concomitantly with prerequisites of creating corpora, the authors divide the corpus building scenario into three main strands including the general corpus, language documentation corpus and research corpus. General corpora are chiefly large in nature and represent language with a wide variety of text types. On the other hand, language documentation corpora are inherently small and limited to specific text excerpts. The last type, that is to say, research corpus, is designed in line with the particular needs of the researcher. Astonishingly, contrary to all conspicuous criteria for corpus building (see e.g. Vian & O'Boyle, 2022), authors purport that "authenticity is not a demarcation criterion for corpora"(p. 95). To put it simply, authenticity means that corpus data are produced and compiled in a natural context without the researcher's intervention (Baker et al., 2006).

Annotation is a process that adds value to a corpus query and entails adding extra data to the corpus (Zufferey, 2020). The authors have dedicated the whole Chap. 7 to this concept to explicate types of annotation such as phonetic and prosodic, morphological, part-of-speech (POS) and discourse annotations. In conjunction with different annotations, a set of pre-eminent corpus annotations and tag sets in linguistic typology such as SCOPIC, Universal dependencies and Multi-CAST are delineated.

A great part of any corpus-based enquiry consists of quantitative analysis, which necessitates a considerable knowledge of statistical analysis. Woven through miscellaneous techniques and formulas, Chap. 8 is exclusively dedicated to statistics in corpus queries. While there are, as the authors indicate, three different sampling

methods such as random, representative and convenience samplings (p.137), they fail to provide the reader with even a rudimentary definition of each. The same goes for the concept of Chi-square test. Indeed, this is expounded without giving the reader clues as to what kind of data (nonparametric) is required for running this test. In line with the statistical analysis, the authors clarify various types of variables, that is to say, dependent and independent ones as two key concepts in research methods in Corpus Linguistics. The same goes for different types of multivariate predictive approaches such as mixed and non-mixed effects regression. These too are offered without definition.

Following the statistical perspective in Chap. 8, the next chapter is a practical one, which falls within the scope of corpus-based studies. While it is not evident why from among a wide range of topics the authors have selected Sociolinguistics as a case in point, they investigate this sub-branch of Linguistics through the lens of Corpus Linguistics. The authors shed light on demographic data about participants. As the authors purport, large corpora can be utilized for investigating dialect varieties as well as ethnicity-conditioned variations in English.

In Chap. 10, emphasis is put on language documentation and its divergence with Corpus Linguistics. As the authors explicate, language documentation is "relatively stable, that is to say the primary data and related metadata are not further manipulated" (p.190), whereas corpus compilation is, by nature, dynamic, meaning that data can be added to or deleted. Barth and Schnell explain how open-ended collections which are extracted from language documentation projects can be utilized as corpora for doing linguistic research. They also explain how larger-scale multilingual corpora are compiled based on language-documentation corpora in order to investigate human languages.

The last chapter of this book centers around the concept of corpus-based typology, which comes to offer "worldwide linguistic diversity (…) to compare languages" (p.197). Putting aside two customary approaches of linguistic typology, that is to say, traditional and multivariate typologies, Barth and Schnell resort to a more recent one called corpus-based typology as a method to delve into themes of universal patterns of language as well as language particularities, which are unique to each language. In addition, they believe corpus-based typology to be an appropriate approach for extracting information on themes of universal features of language as well as diversified and universal features of languages.

There is a misunderstanding with regard to the introduction and definition of corpora. As a case in point, it is not clear why there is no definition of comparable corpora; whereas other types of corpora are clearly explicated. In the same vein, it is unclear why the authors have categorized comparable corpora as a subcategory of multilingual corpora, as comparable corpora can be either monolingual or multilingual (Dash & Arulmozi, 2018). In the same fashion, while the reader is confronted with the practical aspects of deploying a corpus-based methodology in miscellaneous areas with vivid examples in Chap. 4, it is only in Chap. 5 that the reader is equipped with the basic corpus-based queries and techniques. It is not clear why the readers have to encounter authentic applications of corpora before they have been introduced to corpus queries. In addition, there is a big thread of literature on corpora and their versatility for translation research (Vahseghani Farahani, 2022). However, the

authors did not dedicate any attention to corpus-based translation studies and creating parallel corpora in this book.

It is no exaggeration that these shortcomings will not detract from the strengths of this work of scholarship. This book can have pedagogical implications for a wide range of researchers running the gamut from linguists, language teachers to translation studies readers. In addition, kudos must be awarded to the high quality of the print, which makes this book into an enjoyful one to read. Moreover, this practical book successfully holds an equilibrium between theory and practice in Corpus Linguistics.

It is axiomatic that this praiseworthy book strives to add to the knowledge of the competent reader. Notwithstanding, despite the claim of the authors that this book is an introduction to corpus linguistics, this book is of little, if not any, value for the novice and newcomers in the area of Corpus Linguistics as it is not regarded as an introductory guide for them. This value-laden book behooves only the seasoned scholars and researchers in the realm of Corpus Linguistics.

**Acknowledgements** The author of the reviewed book is acknowledged for their insightful and thought-provoking contribution to the field. Gratitude is expressed to the publisher for providing a copy of the book for review. The article 's development is attributed to the invaluable feedback and support of the editor and colleagues.

**Authors' contributions** Dr. Farahani had the idea for the article, both authors performed the literature search and Dr. Ghane drafted and/or critically revised the work.

**Funding** No.
Open Access funding enabled and organized by Projekt DEAL.

**Data Availability** No.

## Declarations

**Ethics approval and consent to participate** Approved.

**Consent for publication** Yes.

**Competing interests** No.

## References

Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburg: Edinburg University Press.
Dash, N. S., & Arulmozi, S. (2018). *History, features, and typology of language corpora*. Berlin: Springer.
Dash, N. S., & Ramamoorthy, L. (2019). *Utility and application of language corpora*. Berlin: Springer.
Paquot, M., & Gries, S. T. (2020). *A practical handbook of corpus linguistics*. Berlin: Springer.
Sekhar Dash, N., & Arulmozi, S. (2018). *History, features, and typology of language corpora*. Berlin: Springer.
Seoane, E., & Biber, D. (2021). *Corpus-based approaches to register variation*. Amsterdam: John.
Vasheghani Farahani, M. (2022). *Writer-reader interaction by Metadiscourse features: English-Persian translation in legal and political texts*. Berlin: Frank & Timme.
Viana, V., & O'boyle, A. (2022). *Corpus linguistics for English for Academic Purposes.* New York: Routledge.
Zufferey, S. (2020). *Introduction to corpus linguistics*. New Jersey: John Wiley & Sons.