**FACTA UNIVERSITATIS**
Series: **Linguistics and Literature** Vol. 21, Nº 1, 2023, pp. 151 - 153

**Book Review**

# Anatol Stefanowitsch
# CORPUS LINGUISTICS: A GUIDE TO THE METHODOLOGY

Berlin: Language Science Press, 2020

## Vladan Pavlović

Faculty of Philosophy, University of Niš, Serbia

*Corpus linguistics: A guide to the methodology* is a relatively recent monograph of Anatol Stefanowitsch, currently a Full Professor in the area of the structure of modern English at the Interdisciplinary Center for European Languages of the Free University of Berlin. He is also an exceptionally influential linguist, whose publications, according to his Google Scholar profile[1], have been cited over 8000 times.

The given book, in the words of its author, is an introductory textbook in the area and focuses on methodological issues – on how to approach the study of language based on usage data and what problems to expect and circumvent (Stefanowitsch, 2020: vii). As such, it has been included in the recommended literature in the courses entitled *Methodology of scientific research* (MA level) and *Methodology of linguistic research* (PhD level) at the Department of English, Faculty of Philosophy, University of Niš.

The book consists of 14 chapters, which can be organized into three groups.

The first group, encompassing chapters 1 through 6, introduces the fundamentals of corpora, of corpus linguistics, of types of data and of significance testing. Chapter 1 presents the arguments *pro* and *contra* corpus data, the issue of intuition data *vs.* corpus data, and the issue of corpus data in subdisciplines of linguistics such as language acquisition, historical linguistics, psycholinguistics, dialectology and sociolinguistics, conversation and discourse analysis.

Chapter 2 provides definitions of the linguistic corpus, with the focus on authenticity, representativeness, size and annotations. The notion of corpus linguistics is defined in several consecutive steps, until eventually arriving at the definition according to which it represents *the investigation of linguistic research questions framed in terms of the conditional distribution of linguistic phenomena in a linguistic corpus* (*ibid*., 56). Having in mind such a definition, in Chapter 3 the author addresses the logic and practice of formulating research questions and testing scientific hypotheses, considers the operationalization of definitions in corpus linguistics and provides general remarks on the research cycle, i.e. on the place of

Submitted March 17, 2023; Accepted April 13, 2023
**Corresponding author**: Vladan Pavlović
University of Niš, Faculty of Philosophy, Ćirila i Metodija 2, 18101 Niš, Serbia
E-mail: vladanp2@gmail.com
[1] https://scholar.google.com/citations?user=CmXWVooAAAAJ&hl=en

hypothesis testing in scientific research practice, especially in view of the well-known points made by Karl Popper, one of the most influential philosophers of science of the 20th century.

Chapter 4 addresses data retrieval – the ways of searching a corpus for various linguistic phenomena, and automatic and *ad hoc* annotation of corpus data.

Chapter 5 discusses types of data and the relevant descriptive statistics applicable to them. In other words, this chapter discusses nominal data alongside percentages and observed and expected frequencies, ordinal data and the related concepts of the medians and frequency lists and modes, and cardinal data and the related concept of means.

Chapter 6 addresses significance testing, including statistical (null) hypothesis testing and the various statistical tests applicable to each of the given types of data. In that sense, this chapter presents the logic of the use of the chi-square test with nominal data (in two-by-two and one-by-n designs), of the Mann-Whitney *U*-test with ordinal data, and of Welch's *t*-test (and the normal distribution requirement) for cardinal data. It thus also enables the readers to get an insight into what goes on "under the hood" (Egbert, Larsson, and Biber 2020, 2), i.e. into the logic behind the use of various statistical tests applied to language data.

The second part of the book, encompassing chapters 7 through 11, presents case studies of the application of corpus linguistics as a scientific method in various domains. Chapter 7 addresses the domain of collocations. It presents collocations as a quantitative phenomenon and discusses methodological issues in collocation research, including measuring the strength of association between two or more words, i.e. the effect sizes of their co-occurrence, with the help of the chi-square, the mutual information, the log-likelihood ratio, the minimum sensitivity and Fisher exact tests, and presents a comparison of the various measures of association strength obtained by of such tests.

Chapter 8 focuses on the study of various syntactic structures with respect to their semantic, informational-structural and other restrictions placed on their particular slots, and their distribution across language varieties and texts. This chapter also heavily relies on collostructional analysis, a statistical analysis that measures the degree of attraction or repulsion that words exhibit to syntactic constructions, pioneered by Stefanowitsch himself and Stefan Thomas Gries. Some of the syntactic constructions explored from the given perspective in this chapter are the ditransitive (*vs.* the prepositional dative) construction, the complementation of the verbs *begin* and *start*, and the *to*- *vs*. *that*-complements. Chapter 9 addresses the issue of quantifying morphological phenomena (such as types, tokens, and hapax legomena) and presents several case studies related to morphemes and stems (such as the phonological constraints on *-ify*, the semantic and phonological differences between *-ic* and *-ical*, and affix combinations). It also provides case studies addressing the relation between productivity of derivational morphemes, on the one hand, and speaker sex and genre, on the other.

Chapter 10 addresses text analysis, including keyword analysis and case studies in: a) language varieties (such as academic English, b) comparing speech communities in an attempt to identify cultural differences between them, c) the co-occurrence of lexical items and demographic categories, and d) the relation between texts, on the one hand, and ideology and time periods, on the other hand. Such analyses testify to the importance of the corpus linguistic methods in sociolinguistic research. Finally, chapter 11 addresses the corpus linguistic exploration of metaphor starting both from the source and the target domains. The case studies starting from the former domain address lexical relations and word forms in metaphorical mappings, while those in the latter address happiness across cultures and intensity of emotions. This is followed by case studies in the relation between metaphor and text (such

as identifying potential source domains, metaphoricity signals and metaphor and ideology) and a case study in metonymic expressions (such as subjects of the verb *bomb*). Furthermore, this chapter also provides some practical advice on how to overcome obstacles imposed by the fact that metaphor and metonymy are linguistic phenomena not purely lexical in their nature.

The third part of the book consists of an Epilogue (Chapter 12), Study notes presenting resources and further reading for each of the preceding chapters (Chapter 13), and Statistical tables with critical values for the results obtained by the various statistical tests listed above. This part of the book also presents References with about 300 sources as well as the Name index and the Subject index. The book also offers supplementary materials online via the relevant links.

This book represents a valuable source for students and others interested in corpus linguistics, and an excellent starting point for delving further into the area, such as getting acquainted with various systems for statistical computation and graphics, such as R (R Core Team 2019), or with the (statistical) corpus tools, such as *AntConc* (Anthony, 2022), *CQPweb* (Hardie 2012), *LancsBox* (Brezina, Weill-Tessier, and McEnery 2020), or *WordSmith* (Scott 2022).

**Acknowledgement**: *Prepared as a part of the project Scientific publications in teaching English Linguistics and Anglo-American Literature and Culture, conducted at the University of Niš – Faculty of Philosophy (No. 300/1-14-1-01).*

## REFERENCES

Anthony, Lawrence. 2022. AntConc (Version 4.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Brezina, V., Pierre Weill-Tessier, and Anthony McEnery. 2020. #LancsBox v. 5.x. [software]. Available at: http://corpora.lancs.ac.uk/lancsbox.

Egbert, Jesse, Tove Larsson, and Douglas Biber. 2020. *Doing linguistics with a corpus: Methodological considerations for the everyday user*. Cambridge: Cambridge University Press.

Hardie, Andrew. (2012). "CQPweb—combining power, flexibility and usability in a corpus analysis tool." *International Journal of Corpus Linguistics*, 17(3), 380-409.

R Core Team. 2021. R: A language and environment for statistical computing (Version 4.1.2) [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Scott, Mike. 2022, WordSmith Tools. Version 8 (64-bit version). Stroud: Lexical Analysis Software.