https://doi.org/10.33919/JCAL.23.1.6

How Much Linguistics in Corpus Linguistics? Review of *Doing Linguistics with a Corpus* by Egbert, J., Larsson, T. and Biber, D. (2020). Cambridge University Press.

Maria Stambolieva

New Bulgarian University

The publication contains 80 pages (with References), organized in eight chapters, and an Appendix. In their abstract to the text, the authors define their work as an attempt to marry traditional corpus linguistics, with its carefully designed and minutely analysed texts, to the modern state of the art – marked by digitalization, abundance of texts and text collections, and wide array of tools. The stated goal is "to explore ways (…) to improve how we approach linguistic research questions with quantitative corpus data".

The **introduction** begins with a provocative parallel between quantitative linguistics and car driving, followed by a quick review of the chapters to come. The parallel with driving stands on the observation that, with recent technological advances, it is increasingly easy to drive a car without knowing much about the engine – just as it has become easy, in corpus linguistics, to use readily available corpora and corpus analysis tools to answer research questions or to obtain results. And just as some understanding of how the car works can be useful in cases of malfunction, the authors insist that basic knowledge of linguistics: an understanding of the nature of a corpus, the linguistic characteristics of the data or the ability to interpret quantitative results are necessary for corpus linguistic analysis. Linguistic skills are involved in the formulation of linguistic research questions and in the interpreting of quantitative results as linguistic patterns. In all the following chapters of the book, this point is illustrated with relevant case studies and emphasized with key points and key considerations.

Getting to Know Your Corpus (Chapter 2) takes up the long-standing, Sinclair vs Biber, discussion on corpus makeup. Attention must of course be paid to both corpus composition and corpus size and, all things being equal, a bigger corpus is an advantage. The reader is nevertheless warned that all things are almost never equal, and decisions on the composition of the corpus should not be taken lightly. Corpus linguists are not, as a rule, interested in how language is used in a corpus as such, but in how language is used in a target register, dialect, etc. – hence the importance of representativeness in corpus design. For the decision process, the authors recommend the following: 1/ careful examination of the metadata and documentation; 2/ examination of the actual texts. The requirement for careful examination of the metadata and documentation before using a corpus for specific research is well supported by the results of a Case study: an investigation of the use of nominalisations and linking adverbials in the target domain of published academic writing, as represented in two subcorpora: the academic sub-corpus of the British National Corpus (BNC_AC) and the academic subcorpus of the Corpus of Contemporary American English (COCA_AC).

The third chapter, **Research Designs: Linguistically Meaningful Research Questions, Observational Units, Variables, and Dispersion,** is a presentation of several topics required to understand how quantitative corpus analysis relates to tangible linguistic descriptions. The two underlying major concepts here are research design and research questions. "Research design" is defined as the way in which quantitative linguistic data is collected and organized. Research questions specify what we want to learn about language use by doing corpus analysis; accordingly, these questions dictate the research design. Conversely, once data has been collected according to a particular research design, it should only be used to answer certain types of linguistic research questions. The importance of research design is exemplified with the investigation of research questions involving dispersion, and supported with a case study on English genitives in a variationist, whole-Corpus, and text-linguistic research. The chapter concludes with the following key considerations: 1/ observational units can be defined at the level of the linguistic feature, the text, or the corpus; 2/ results from a variationist research design have a dramatically different interpretation from those from descriptive linguistic research designs; 3/ the text-linguistic research design has many advantages over the whole-corpus research design.

Chapter 4, **Linguistically Interpretable Variables,** addresses the need to ensure that all variables used in a corpus study are linguistically interpretable. A linguistic variable is interpretable when its scale and values represent a real-world language phenomenon that can be understood and explained. To illustrate the points made in this section, the authors present two short case studies: "Measures of collocation" (Case study 1) and "The linguistic interpretation of "keyness" measures" (Case study 2). Case Study 1 explores the use of concordancing for one of the primary goals of the study of collocation – the study of the extended meanings of words beyond their traditional dictionary definitions. A very clear example is presented: the verb *to cause*, traditionally defined as "make something happen". Corpus research demonstrates that this verb frequently co-occurs with words referring to negative events – hence the extended meaning of the verb: "make something *bad* happen". Another example is an exploration, based on immediate context, of the way *man* and *woman* are characterized in the corpus COCA_AC. In summary, the simple frequency approach to collocation is argued to be more appropriate for the purpose of discourse characterization than statistical collocational measures, as the two produce different results and require different linguistic interpretations. Case Study 2 is a presentation, following Egbert and Biber (2019), of keyword analysis and "text dispersion keyness". Text dispersion keyness is argued to have two major advantages: (1) it takes into account the dispersion of a word across the texts of a corpus and (2) it is more directly interpretable in linguistic terms than traditional measures – because a text is a valid unit of language production, while a corpus is not.

Chapter 5, **Software Tools and Linguistic Interpretability**, presents a central thesis of this work, based on a case study analysis of grammatical complexity measurement – complex nominals. The measure of complexity of nominals is problematic because, among other things, it does not distinguish between pre- and post- modification and between single and multiple modification. The authors conclude that in order to ensure reliable conclusions based on existing corpus-analysis tools, considerable post-processing is needed – involving, for instance, the evaluation of accuracy. The analysis of a number of smaller corpora, while more time and work consuming, yields results that are more accurate and linguistically meaningful and interpretable. Researchers are advised to choose or develop such tools and measures that are linguistically sound and well documented.

The question of what constitutes appropriate statistical methods is the focus of Chapter 6, **The Role of Statistical Analysis in Linguistic Descriptions.** Following examination of Null hypothesis significance testing (NHST) as a statistical paradigm, the authors (while not denying the usefulness of statistical methods) here again stress the importance of staying close to the language data. Language "is, and should remain, the primary focus of corpus linguistic investigations". Sophisticated statistical methods often create layers of distance between corpus researchers and the language data they aim to describe, which could affect negatively the linguistic validity of the results. Put differently, any kind of abstracting away from the language data increases the risk of obtaining linguistically uninterpretable results – which, in turn, is more likely to lead to misinterpretations and unsatisfactory conclusions. The chapter ends with the following key considerations: 1/ because sophisticated statistical methods often force researchers to abstract away very far from the language data, it is important to employ minimally sufficient statistical methods and remain as close as possible to the language data; 2/ NHST should always be complemented by consideration of descriptive statics and effect sizes; 3/ in order to interpret numeric results, conscious effort should be made to return to the language data.

Chapter 7, **Interpreting Quantitative Results**, can be seen as a summary and generalization of the issues discussed in the previous chapters. The authors argue that computational linguistics is still linguistics, and that linguistics is done by linguists. Computers can of course process corpus data, but they cannot interpret them as "meaningful patterns of language use". The following sources for qualitative interpretation of data that linguists rely on are highlighted: (1) linguistic context, (2) text-external context (above all, metadata), and (3) linguistic principles and theories. Usage-based linguistics, which "explores how we learn language from our experience of language" (Ellis, 2019), is quoted as a "good example of a healthy relationship between linguistic theory and quantitative corpus linguistics". The key takeaways from this chapter are: 1/ linguistics is done by linguists, not by computers; 2/ in order to be useful, quantitative corpus linguistic analysis should be coupled with sound qualitative interpretation; 3/ in their interpretation of quantitative corpus findings, researchers should be guided by linguistic context, text-external context and linguistic theory.

In the final chapter **(Wrapping Up)** the authors summarise the motives which led them to writing the book: reinstating linguistics at the center stage of (quantitative) corpus linguistic research and pointing to means to achieve this. The output of quantitative analysis is data. Data "are to information what iron ore is to iron: nothing can be done with data until they are processed into information". Information is contained in descriptions, answers to questions that begin with such words as *who*, *where*, *when*, and *how many*. In other words, "[i]nformation is born when data are interpreted" (Stallings, 1989, 2). Statistical analysis can provide us with data, but that data must be interpreted if it is to be useful for linguistic description. Linguistic research begins with the formulation of meaningful linguistic research questions and the purpose of corpus design is to answer these questions.

Concluding remarks

The book reviewed focuses on important issues related to the role of linguists, linguistic theory and linguistic research questions in modern corpus linguistics – issues which have been by-passed or ignored for some time, and particularly in the last decade. Backed by clear argumentation and illustrated with ample data, this Element – as the authors have chosen to define their text – manages to cover substantial ground against prevailing winds and currents. For the linguists in the profession, it is a godsend. For other researchers in the field, it is a must-read.

References

Egbert, J. and Biber, D. (2019). Incorporating Text Dispersion into Keyword Analyses. *Corpora*, 14(1), 77–104.

Ellis, N. (2019). Usage-based Theories of Construction Grammar: Triangulating Corpus Linguistics and Psycholinguistics. In: Egbert, J. and Baker, P., (eds.). *Using Corpus Methods to Triangulate Linguistic Analysis*. New York: Routledge.

Stallings, W. (1989). *Data and Computer Communications.* 4th ed. New York: Macmillan.