RiCL Research in Corpus Linguistics

Review of Pérez Paredes, Pascual. 2020. *Corpus Linguistics for Education: A Guide for Research*. London: Routledge. ISBN: 978-0-367-19843-5. https://doi.org/10.4324/9780429243615

Barry Pennock-Speck
University of València / Spain

This volume is the eighth in the series, *Routledge Corpus Linguistics Guide*, edited by Michael McCarthy and Anne O'Keeffe. In the description of the book, we are told that it "provides a practical and comprehensive introduction to the use of corpus research-methods in the field of education" and that the approach is "hands on."[1] After a critical reading of the volume, I can state without hesitation that both assertions are true. *Corpus Linguistics for Education: A Guide for Research* is targeted at both students and researchers who are studying or carrying out research in Corpus Linguistics in the field of education. There are eight chapters including an introduction and a short conclusion. No previous knowledge of Corpus Linguistics is assumed as this volume in the series is explicitly described as an introductory textbook. The style is unpretentious and great efforts are made to make the volume readable. For example, Chapter 1 starts off in a light-hearted manner with "If you expect to find a definition of Corpus Linguistics in this opening paragraph, you will not be disappointed" (p. 1) and the author even makes a mouthful, such as "the typicality of hegemonic discourse" (Baker 2006), palatable by offering interesting quotes that demonstrate the usefulness of the term when related to disability and the press's depiction of Islam and Muslims (p. 13). The 42 figures and 51 tables are designed to guide the reader through what is sometimes quite complex material. In Figure 1.4, for instance, the reader is shown how to differentiate a corpus as primary data from one used as secondary data. It is also used to further explain these two approaches and is then followed by further explanations. Tables are also often used

[1] https://www.routledge.com/Corpus-Linguistics-for-Education-A-Guide-for-Research/Perez-Paredes/p/book/9780367198435

in an explanatory fashion. For example, Table 1.3 shows the differences between positivism and phenomenology, the approach taken by researchers in each paradigm and the methods they use. Each chapter ends with a notes section made up of notes proper and the addresses of web sites, and a list of references. Both the links and the references are very useful for those who might want to look into aspects dealt with in each chapter in more depth. Competences are considered as well as knowledge through the 18 skills introduced chapter by chapter. In Chapter 4, skills one to 11 are reviewed, and skills 12 to 17 in Chapter 7. The book is complemented by a link to a document that includes suggested answers to the skill review questions.

The philosophy of Corpus Linguistics is summed up in the first paragraphs of Chapter 1 as the study of the language of real life and the empirical analysis of attested usage of actual language. The author then goes on to provide a definition of corpus, "a large body of texts" (p. 1), and further on describes a corpus as being both an instrument and a method designed to answer research questions. From an epistemological point of view, Corpus Linguistics is situated within the scientific paradigm, as it uses mainly quantitative methods and large samples, that is, corpora. There follows a description of several corpora and the research questions they helped to answer. The author outlines several important concepts in Corpus Linguistics such as accountability, falsifiability, replicability and representativeness using examples and quotes from leading authors in the field.

Chapter 2, the shortest of the chapters, focuses on text analysis in the field of education research. In the first part of the chapter, the author gives an account of the two main qualitative approaches employed in text analysis in the field, content analysis and theme analysis. He subsequently goes on to provide a brief description of software packages such as *NVivo*,[2] *MAXQDA*[3] and *AntConc* (Anthony 2019) that can be used to code texts. We are then offered an overview of Conversation Analysis and Discourse Analysis —approaches that have used Corpus Linguistic tools to achieve their objectives (cf. Flowerdew 2012 or Walsh 2016). It is when we reach Section 2.2.1 that we are introduced to the first dissensions that exist in Corpus Linguistics, for example, corpus-based vs. corpus driven approaches and theory-driven and data-driven approaches. It is also after this section that we are introduced to quantitative and

[2] http://image-analysis.com/
[3] https://www.maxqda.com/

qualitative approaches used alongside Corpus Linguistics. In the final section of the chapter, the focus is on the register perspective found in the work of Biber and Conrad (2009).

In Chapter 3, we are introduced to Corpus Linguistic approaches applied to the study of language use. Here the author underlines the importance of Corpus Linguistics to discover regularities and patterns in texts, which can help us to understand better the textual habits of communities. The author admits that, to a certain extent, Corpus Linguistics reduces the complexity of a text to a simpler form to make it more comprehensible. For the first time in the volume, we come across case studies. In the first study, we are introduced to both qualitative and Corpus Linguistic methods to analyse interviews, in the second to content analysis and Corpus Linguistics to examine educational policies. The case studies are useful in that, through practical examples, we are shown the differences and affordances of Corpus Linguistics when compared to qualitative approaches that employ interviews and thematic analysis mentioned in an earlier chapter. Through copious examples and explanations, in Section 3.2, we are given our first glimpse into the practicalities of Corpus Linguistics in the shape of concordance lines. The author includes a practical six-step procedure to analyse concordance lines. Technical terms, such as 'lemmas', 'types', 'tokens', 'nodes', are introduced at intervals, which, together with the screenshots from *AntConc*, *WordSmith* (Scott 2020) and *Sketch Engine* (Kilgarriff *et al.* 2014), makes it easier to grasp their meaning. The reader is also shown how the results from word lists can be exported. The handling of frequencies is explored next. Emphasis is placed on the importance of considering the size of corpora such as the *British National Corpus* (BNC; BNC Consortium 2007) or the *Corpus of Contemporary American English* (COCA; Davies 2008–) and the need to calculate a term's relative frequency. As in the case of concordance lines and word lists, the reader is shown not only what a lemma is but how a lemma list is loaded into *AntConc* or where to go to download a lemma list (Mike Scott's web site).[4] Finally, definitions of *collocation* and *collocate* are provided and their importance for Corpus Linguistics is explained. Once more, figures and tables make it easier for the reader to understand the concepts that have been introduced.

[4] https://www.lexically.net/

Chapter 4 focuses on designing corpora. The first section covers corpus size and data collection. The readers are warned against attempting to compile unrealistically large corpora. A corpus should be large enough to be representative of the type of language being studied and, importantly, while taking into account the time needed to compile it (Reppen 2010). The reader is then provided with two lengthy case studies. The first uses Corpus Linguistics as the main research methodology to analyse issues involving early childhood education in Australia. The reader is shown the research questions that were drawn up and the analyses performed to answer them. The second case study involves embedded Corpus Linguistic research methods. However, the reader is not given a definition of the meaning of 'embedded Corpus Linguistics'. This study involves 'narrative policy analysis' and Corpus Linguistic methods to analyse aspects of literacy education in Canada after the crisis in 2008.

The second section, 4.2, goes into the basics of comparing corpora and significance testing but I have only found an indirect reference to significance testing in the reference to keyword analysis. The author puts forward that Corpus Linguistic research is frequently comparative and that, when two corpora are analysed, it is normal to gauge the differences in usage between two (or more) corpora or use a second one as a reference corpus to carry out a keyword analysis. To illustrate the comparison of corpora, the author looks at educational policy publications from the UK and New Zealand. Section 4.2.1 examines the functionality of Part-of-Speech (POS) tagging. This, the author states, adds sophistication to the searches as POS tagging can be combined with words or lemmas. The UK and New Zealand corpora are used to illustrate POS tagging. The numerous figures used to show POS tags in *Sketch Engine* and *AntConc* are extremely useful. Information is also given on freely available POS taggers that are freely available. The final sections of Chapter 4 are given over to a review of skills one to 11.

Chapter 5 is dedicated to describing the transcription and annotation of interview data. The author highlights the labour-intensive nature of transcribing interviews. Transcriptions, we are told, may just contain what was being said but could also include other details such as the tone of voice, inflection, emphasis, pauses, interruptions, etc. We are given a glimpse into the many decisions that need to be made when transcribing. The *Child Language Data Exchange System* (CHILDES)[5] is offered as an example of

[5] https://childes.talkbank.org/

thorough transcription and once more, as in earlier chapters, coding is brought up. In Section 5.2, basic transcription techniques are described. Two desktop solutions, *Inscribe*[6] and *EXMARaLDA Partitur Editor*,[7] designed to help researchers with transcriptions, are briefly outlined. The author adds that these can be complemented with software such as *Praat* (Hirst 2013), ELAN (Wittenburg *et al.* 2006) or the *UAM Corpus Tool*.[8] Table 5.1 includes a comprehensive insight into the LINDSEI transcription guidelines.[9] Readers are advised to employ setting brackets to tag annotations such as <foreign> and </foreign> to be able to find annotated text in software such as *AntConc*. Figure 5.2 is provided to illustrate how this is done. The final Section 5.3 emphasises the need to add structure and metadata to a corpus. 5.3.1 explains how to annotate a corpus with one's own tags to get the most out of searches using *Sketch Engine*. The final section, 5.3.2, deals with annotation using standard XML guidelines. The author suggests a *Text Encoding Initiative* (TEI) template to gather metadata (Pérez-Paredes and Alcaraz-Calero 2009) and goes into great detail in its description. Chapter 5 is not the longest chapter in the volume, that honour goes to Chapter 6, but to my mind it is the most complex. Nevertheless, great care is taken to make the explanations as clear and comprehensive as possible.

The remit of Chapter 6 is to provide the reader with insights into the Corpus Linguistic analysis of vocabulary. The keyword analysis of a corpus of peace treaties is employed to show how the concept of education is used in the texts. Keywords, as the author explains, can help to highlight the words that characterise the corpus under scrutiny. The reader is provided with a lengthy but very helpful step-by-step guide to keyword analysis using the corpus of peace treaties. In Section 6.3 there is a guide to researching nouns and noun phrases focusing on the examination of their colligational behaviour using *Sketch Engine*, first by focusing on individual nouns, in this case *education*, and then multiword units. The final section, 6.4, on the lexicon of children's literature involves various corpora. It is here that we come across N-Grams, which are described in depth for the first time. Strangely, N-Grams are not highlighted as essential terminology as are other terms, but Table 6.11 offers a summary of how they are used.

[6] https://www.inqscribe.com/
[7] https://exmaralda.org/en/partitur-editor-en/
[8] http://www.corpustool.com/index.html
[9] https://uclouvain.be/en/research-institutes/ilc/cecl/transcription-guidelines.html

Chapter 7 centres on examining talk and how to tease out the collocations and patterns in conversations and interviews. Corpus Linguistics, the author states, gives researchers unique insights into how language is used. Using the *Backbone Corpus of English as a Lingua Franca* (Kohn 2012), readers are shown how Corpus Linguistic methods can serve to carry out more complex searches. There follows a review of the major differences between spoken and written language. The systemic functional grammar approach is used (Locke 2004), for example, to classify transitivity. Section 7.2 outlines how to do multiword keyword searches through Corpus Query Language (CQL) in *Sketch Engine.* In Section 7.2.3 readers are instructed on how to run a search that will show how family life is impacted by work in the *Backbone Corpus*. This is possible as the coders had included the information in the corpus. The author links the findings to the principles talked about in Chapter 1. Section 7.3 reviews skills 12 to 17, that is, those found in Chapters 5 to 7.

Chapter 8, the conclusion, is short but emphasizes the positivist nature of Corpus Linguistics and finishes with skill 18, that is, remaining critical. Here the reader is warned to be aware that what they do with Corpus Linguistics depends on the design of the corpus, the methods used and the interpretation of the results.

All in all, *Corpus Linguistics for Education: A Guide for Research* embodies a model introductory textbook. Very difficult concepts are introduced and fleshed out in a logical and straightforward manner. More importantly, the theory is linked, at all times, to the actual practice of Corpus Linguistics. This process is helped along greatly by the generous offerings of figures, tables, links to websites and copious references to seminal publications. As the readers are shown the workings of some of the most popular Corpus Linguistic programmes, and are given the links to several freely downloadable corpora, I imagine that several would be sorely tempted to try their hand at Corpus Linguistic analysis even before finishing the book—as I did.

REFERENCES

Anthony, Laurence. 2019. *Antconc* (version 3.5.8). Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/antconc/

Baker, Paul. 2006. *Using Corpora in Discourse Analysis.* London: Continuum.

Biber, Douglas and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

BNC Consortium. 2007. *The British National Corpus*, version 3 (BNC XML Edition). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/

Davies, Mark. 2008–. *Corpus of Contemporary American English*. https://www.english-corpora.org

Flowerdew, Lynne. 2012. Corpus-based discourse analysis. In James Gee and Michael Handford eds. *The Routledge Handbook of Discourse Analysis.* London: Routledge, 174–187.

Hirst, Daniel. 2013. Anonymising long sounds for prosodic research. In Brigitte Bigi and Daniel Hirst eds. *Tools and Resources for the Analysis of Speech Prosody*. Aix-en-Provence: Laboratoire Parole et Langage, 36–37.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1/1: 7–36.

Kohn, Kurt. 2012. Pedagogic corpora for content and language integrated learning. Insights from the BACKBONE Project. *The Eurocall Review* 20/2: 3–22.

Locke, Terry. 2004. *Critical Discourse Analysis.* London: Bloomsbury.

Pérez-Paredes, Pascual and José M. Alcaraz-Calero. 2009. Developing annotation solutions for online data driven learning. *ReCALL* 21/1: 55–75.

Reppen, Randi. 2010. Building a corpus: What are the key considerations? In Anne O'Keeffe and Michael McCarthy eds. *The Routledge Handbook of Corpus Linguistics.* London: Routledge, 31–37.

Scott, Michael. 2020. *WordSmith Tools version 8*. Stroud: Lexical Analysis Software.

Walsh, Steve. 2016. Applying corpus linguistics and conversation analysis in the investigation of small group teaching in higher education. In Halina Chodkiewicz, Piotr Steinbrich and Malgorzata Krzeminska-Adamek eds. *Working with Text and Around Text in Foreign Language Environments.* Bern: Springer, 205–222.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk and Daniel Tapias eds. *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, 1556–1559.

Reviewed by
Barry Pennock-Speck
University of València
Faculty of Philology, Translation and Communication
Department of English and German
46010. Valencia
Spain
E-mail: barry.pennock@uv.es