

BOOK REVIEWS

Di Cristofaro, M. (2023). *Corpus approaches to language in social media*. Taylor & Francis. xii + 386 pp.

Reviewed by Aleksandra Sevastianova (University of Edinburgh)

It is often argued that the main benefit of using corpus methods in linguistic research is the ability to accurately and automatically process large volumes of data (e.g. Brookes & McEnery, 2019). Corpus methods are becoming increasingly popular in social media research, where the ever-growing amount of data can be overwhelming when studied manually. In this book, Di Cristofaro explores the connections between corpus linguistics, computer sciences, and digital humanities, promoting interdisciplinarity as the basis for modern social sciences. As reflected in the title, the author uses the term ‘corpus approaches’ (p.2) rather than ‘corpus methods’ or ‘corpus linguistics’ to highlight a broader understanding of the concept and stress the need to incorporate techniques from various fields into social science.

Di Cristofaro provides a practical, exhaustive guide to interdisciplinary corpus research, paying special attention to data collection and corpus design. The technical aspects of corpus research are brought to the fore, making this book stand out from other corpus linguistics guides and handbooks which typically focus on the linguistic aspects (e.g. McEnery & Hardie, 2011; Collins, 2019). Di Cristofaro highlights “the importance of engaging with the code” (p.376), which is exactly what the book does. It contains detailed explanations of the technicalities related to data creation and corpus design as well as extended Python scripts accompanied by comments, which might be useful for those looking to create their own tool to meet the needs of a particular study. Moreover, the book provides a lot of platform-specific information on data collection and processing, making it a valuable resource for social media researchers.

The book is divided into seven parts: an introduction, five main chapters, and a conclusion. In the introduction, Di Cristofaro states the aim of the book: to enhance the corpus approach by adding the technical concepts and procedures necessary for a better understanding of corpus data. This chapter lays the foundation for further discussion through a theoretical reflection on the influence of the digital environment on such social processes as cognition, communication, and cooperation. The author then introduces the fields of digital humanities and corpus linguistics, emphasizing their interconnectivity and transdisciplinarity throughout the remainder of the chapter.

The next chapter explores how social media can be approached as research data. First, the author reflects on the influence of technology on society, cognition, and language. The concept of open source is then introduced, followed by an extended discussion of legal and ethical aspects crucial for social media research. These aspects are sometimes overlooked in corpus linguistics books due to their complexity, specificity, and platform dependence, making this section a useful resource offering the latest information on the issue. Di Cristofaro introduces the main characteristics of a corpus, paying special attention to the concept of representativeness which sparks a lot of discussion in the corpus linguistics community. The author considers such notions as metadata, textual markup, and annotation, suggesting a framework to evaluate social media metadata.

Chapter 3 provides an overview of basic corpus tools and functions. The author lists the most popular corpus tools, defines fundamental concepts of corpus linguistics such as ‘type’, ‘token’, and ‘lemma’, and explains standard functions of these tools including frequency lists, dispersion, concordances, collocations, keywords, and stoplists. While some corpus handbooks move from quantitative to qualitative functions, starting with keywords, then collocations, and finally concordances (e.g. Paquot & Gries, 2021), Di Cristofaro introduces the functions in a different order. This decision reflects the book’s aim of enhancing the reader’s understanding of the technical side of corpus approaches rather than providing an analytical framework for linguistic analysis. The author introduces several methods of analysis that are more characteristic of digital humanities than corpus linguistics such as sentiment analysis and topic modelling, reinforcing the main message of promoting interdisciplinarity and implementation of new methods within corpus linguistics.

In Chapter 4, Di Cristofaro discusses data and the process of corpus design. Throughout the chapter, it is argued that the way data is imagined and constructed is as essential for analysis as the research questions themselves, and that computer sciences and digital humanities are important contributors to more thorough corpus design. This is arguably the most technically loaded chapter, introducing such notions as command-line interface, programming languages (i.e. Python), and various technical formats (e.g. CSV, XML, HTML, and JSON). Besides that, Di Cristofaro explains how to preserve, clean, process, and format the data for successful research, emphasising the benefits of “situated software” over “point-and-click tools” (p.103). This chapter might be challenging for linguists who are not familiar with programming and typically rely on ready-made datasets and corpus tools. Acknowledging this, Di Cristofaro encourages social scientists to focus on “the relation between code and the considerations pertaining to the analysis of language, regardless of their coding skills” (p.102). This seems to be a reasonable approach, since the chapter does not aim to teach

Python but rather to show its potential for corpus linguistics. However, the author provides an extensive reading list for those interested into delving deeper into coding.

Chapter 5 addresses various aspects of data collection. Di Cristofaro begins by outlining some limitations associated with using social media platforms for this purpose before going on to consider the differences between crawling and scraping, discuss the use of APIs, and present several data collection tools. The first tool discussed, *#LancsBox* (Brezina et al., 2015; Brezina et al., 2020), is a widely used tool in corpus linguistics, along with *CQPWeb* (Hardie, 2012), *AntConc* (Anthony, 2023), and others. Di Cristofaro focuses on *#LancsBox* because of its recent addition of a crawling and scraping functionality. Rather than limiting the discussion to tools already used by corpus linguists, the author advocates for exploring data collection tools not developed specifically for corpus linguistics, introducing *Archivebox* (<https://archivebox.io/>), *Trafilatura* (Barbarelli, 2021) and *BeautifulSoup* (Richardson, 2019). The next section provides examples of platform-specific data collection tools for popular social media platforms, along with detailed technical explanations. It is worth mentioning that legal and ethical considerations are left to the reader at this point, meaning that the appropriateness of using the data collection tools discussed in this chapter should be critically evaluated by the reader in every case. The author suggests consulting the “Copyright and Ethics” section, which serves as a good starting point for anyone planning to carry out social media research.

Chapter 6 introduces three case studies that demonstrate how the above-mentioned corpus and digital methodologies have been used at the intersection of linguistics, criminology and law. All three projects summarised were conducted 2014–2018 at Swansea University under the direction of Professor Lorenzo-Dus. The first project analysed crypto-drug market forums to define discourse strategies for establishing trust within drug trading communities on the internet (e.g. Lorenzo-Dus & Di Cristofaro, 2018). The second project explored how far-right groups build and promote their collective identity on social media (e.g. Lorenzo-Dus & Nouri, 2021). The third project contributed to the development of computer technologies aimed at protecting children from online groomers by detecting the linguistic and discursive patterns typical of such individuals in internet communication (e.g. Lorenzo-Dus et al., 2020). Every overview contains a brief background and context outline focusing on details of corpus design and data processing that highlight procedures described earlier in the book (e.g. working with HTML files, creating metadata, data extraction, text normalization). For instance, in summarizing the second case study, Di Cristofaro points out that the corpus design was supplemented with information acquired through sentiment analysis and topic modelling, combining a methodology typical of digital humanities

with corpus linguistics, as discussed in Chapter 3. The author also addresses the challenge of using multiple data extraction sources when creating a corpus. He demonstrates the associated techniques implemented in the study by reproducing the main data processing steps. Another methodological aspect discussed in this overview is the transliteration of emojis at the corpus building stage. While emojis are an indispensable part of social media communication, they can present challenges for corpus research due to their multimodal nature. The approaches outlined in this study may offer viable solutions to these challenges. In terms of corpus linguistics methodology, all three studies use standard corpus-assisted discourse analysis procedures, such as keyword, collocation, and concordance analyses. A useful addition to the chapter would be a more thorough explanation of the linguistic analysis process and the results of the research. Still, the author provides multiple, exhaustive references so that keen readers can access details of the studies mentioned.







Finally, the conclusion reinforces the author's message that collaboration across fields is a welcome development of modern science, leading to beneficial outcomes that cannot be achieved within a single academic discipline. The author highlights the ubiquity of social media in society and, consequently, academia, where the same social media data can be explored from many angles. Di Cristofaro advocates for a reevaluation of the interaction between researchers and computers, promoting proficiency in programming languages as a basis for more efficient interdisciplinary cooperation. At the end of the chapter, he reiterates that corpus linguists need better command of digital skills to succeed in studying the growing amount of digital data. This is an important claim that might capture the direction of the further development of corpus linguistics. Many corpus linguists who prefer to use specific tools in their work might at some point feel limited by what the tool functionality can offer, such that acquiring the fundamentals of programming might soon become a necessity for better understanding of corpus data.

Overall, the book is a successful and innovative work, showing the potential of corpus approaches in interdisciplinary fields. Its elegant, easy-to-follow style makes it a great resource for researchers looking to deepen their knowledge of interdisciplinary social media research. It could also serve as a useful read for corpus linguists who work with ready-made tools and would like to acquire some basic understanding of digital technologies useful for linguistic research. The book broadens the horizons of corpus linguistics by suggesting new approaches, tools, and techniques that are not yet commonly used within the field. When combined with traditional approaches to corpus analysis, some of the tools introduced by Di Cristofaro may help social scientists to achieve more comprehensive results when studying the complicated and ever-changing phenomenon of social media.

Funding

Open Access publication of this article was funded through a Transformative Agreement with University of Edinburgh.

References

- Anthony, L. (2023). *AntConc* (Version 4.2.4) [Computer Software]. Waseda University. <https://www.laurenceanthony.net/software>
-  Barbaresi, A. (2021). Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In H. Ji, J. C. Park, & R. Xia (Eds.), *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing: System demonstrations* (pp. 122–131). Association for Computational Linguistics.
-  Brezina, V., McEnery, T. & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139–173.
- Brezina, V., Weill-Tessier, P., & McEnery, T. (2020). *#LancsBox* (Version 5.x.) [Computer Software]. Lancaster University. <https://corpora.lancs.ac.uk/lancsbox>
-  Brookes, G., & McEnery, A. (2019). Corpus linguistics for indexing. *The Indexer: The International Journal of Indexing*, 37(2), 105–124.
-  Collins, L. (2019). *Corpus linguistics for online communication: A guide for research*. Routledge.
-  Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3), 380–409.
-  Lorenzo-Dus, N., & Di Cristofaro, M. (2018). I know this whole market is based on the trust you put in me and I don't take that lightly: Trust, community and discourse in crypto-drug markets. *Discourse & Communication*, 12(6), 608–626.
-  Lorenzo-Dus, N., Kinzel, A., & Di Cristofaro, M. (2020). The communicative modus operandi of online child sexual groomers: Recurring patterns in their language use. *Journal of Pragmatics*, 155, 15–27.
-  Lorenzo-Dus, N., & Nouri, L. (2021). The discourse of the US alt-right online – a case study of the *Traditionalist Worker Party* blog. *Critical Discourse Studies*, 18(4), 410–428.
-  McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Paquot, M., & Gries, S. Th. (Eds.). (2021). *A practical handbook of corpus linguistics*. Springer.
- Richardson, L. (2019, December 19). *Beautiful soup documentation. Release 4.4.0*. <https://readthedocs.org/projects/beautiful-soup-4/downloads/pdf/latest/>

Address for correspondence

Aleksandra Sevastianova
University of Edinburgh
a.sevastianova@sms.ed.ac.uk

Publication history

Date received: 27 February 2024

Date accepted: 21 March 2024

Published online: 19 November 2024