ICAME Journal • 50(1) • 2026 • 99–102
DOI: 10.2478/icame-2026-0008

ICAME Journal

Joan O'Sullivan, Carolina P. Amador-Moreno & Anne Barron. 2025. *Corpus linguistics for sociolinguistics: A guide for research*. London and New York: Routledge. 270 pp. ISBN 978-1-138-80203-2.

Review

Theresa Neumaier

TU Dortmund University, Germany

Sociolinguistics is concerned with language in its social context, more specifically with how social factors are related to linguistic variation. One way of examining this relationship is by looking at language corpora and employing corpus linguistic methods. Yet, despite their "shared interest in empirical analyses of natural data, corpus linguistics and sociolinguistics have traditionally operated independently" (p. 245; see also Biber 2012: 301). It is the goal of the present book, *Corpus linguistics for sociolinguistics: A guide for research*, to show how both fields can complement each other and to "raise awareness of the myriad of research affordances and opportunities" (p. 252) this yields.

The book is an introductory textbook and designed "as a practical guide to corpus sociolinguistics. It does not assume any technical expertise or prior knowledge of or experience with corpora. Nor does it assume any sociolinguistic knowledge" (p. 3). This mission is visible throughout the book: apart from the introduction and conclusion, each chapter addresses a different sociolinguistic topic – age, gender, style, regional variation, language use online, language change, and diachronic variation in the media. Each chapter begins with an accessible introduction to relevant sociolinguistic concepts and then discusses suitable corpora for analysing sociolinguistic variation in the respective field. Detailed descriptions of published research studies are provided to showcase the application of individual corpus linguistic methods. Furthermore, readers are encouraged to critically engage with the concepts and methods presented, as each chapter includes different types of hands-on exercises (described as Listen/Watch and Discuss, Read and Discuss, Stop and Think, Explore Internet Resources, Analyse and Interpret, Experiment with Corpora) and discussion questions. In addition, the book contains an appendix in which some of the figures are reproduced in larger format and five teaching units made available as additional online materials.

With 55 pages, **Chapter 1** constitutes the longest section of the book. It briefly describes the fields of sociolinguistics and corpus analysis and provides some essential historical and conceptual background of sociolinguistics as a research area. The reader is introduced to the three waves of variation research and to central sociolinguistic methods, before corpus linguistics is presented as further means for doing sociolinguistic research – both as a way of collecting data and as a method for data analysis. The latter receives most focus: central tools for corpus linguistic analysis (frequency counts, concordancing, word lists, KWIC, collocations, keyword analysis, and n-grams/cluster analysis) are presented. Readers are carefully guided through using these operations with the help of step-by-step explanations, which are sometimes explained by images. The chapter concludes with an overview of how corpus linguistics can

* *Corresponding author: Theresa Neumaier, E-mail: theresa.neumaier@tu-dortmund.de*

Open Access. © 2026 Author.
This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

be used for sociolinguistic research and of potential difficulties and pitfalls in doing so. The last section warrants special mention: here, O'Sullivan, Amador-Moreno & Barron manage to make their readers aware of the manifold challenges of corpus sociolinguistics without leaving them frustrated; rather, the section ends with a step-by-step description of how to conduct an analysis while being aware of those difficulties.

After the introductory chapter, the following chapters introduce classic sociolinguistic topics from a corpus linguistic perspective. **Chapter 2** focuses on how corpus linguistics can be employed to investigate the influence of gender and age on language use. In Section 2.2, central concepts related to gender are briefly introduced (sex, gender, gender identities, gender roles). This is followed by an overview of central sociolinguistic studies on gender-preferential linguistic features, such as intensifiers, tag questions, or swear words. O'Sullivan, Amador-Moreno & Barron point out correlations between gender and socio-economic status and widen the scope to include cultural differences in gender variation. Section 2.3 addresses age-variation and language change. Again, the section first provides an overview of essential concepts (communal change, generational change, age-grading, lifespan change, retrograde change), before real- and apparent-time studies are introduced as methods for investigating age-related language change. In Section 2.4, concrete examples for corpus linguistic studies from all three waves of variation research – variationist sociolinguistics, ethnographic approaches, and social constructivist approaches – are presented. Although pointing out that "many corpora are not marked up for speaker age and gender" (p. 81), Section 2.5 presents a description of corpora suitable for research in this area. The chapter concludes with a section dedicated to challenges researchers might face when conducting sociolinguistic corpus analyses focusing on gender or age variation, as well as a summary of future developments and desiderata in the field.

Chapter 3 is concerned with using corpora to analyse stylistic variation in fictional discourse. After a general introduction to style as a sociolinguistic concept, Section 3.2 describes how style markers influence readers' perceptions of characters in literary texts and how corpus analytic tools can be employed to make lexical, grammatical, and thematic patterns visible. Apart from that, corpus stylistics examines the use of dialect in fiction. Potential applications are presented in Section 3.3, which also includes a detailed description of a recent study investigating the depiction of Dublin English in Paul Howard's Ross O'Carroll-Kelly novels (Amador-Moreno & Terrazas-Calero 2022). O'Sullivan, Amador-Moreno & Barron show that "literary dialogue can often be taken as a reflection of everyday community language norms" (p. 116) and provide concrete tips for readers wanting to create their own corpus of fictional texts, specifically addressing issues such as representativeness and balance, copyright, data storage, and annotation (Section 3.4). While the clear focus of the chapter is on stylistic variation in literary texts, Section 3.5 briefly addresses applications of corpus stylistics in non-fictional writings (such as policy documents or environmental discourse). Limiting the scope of corpus stylistics to literary texts has been criticised repeatedly (e.g. Wu & Li 2025; McIntyre & Walker 2019), so the inclusion of non-fictional texts in this chapter constitutes a welcome step forward, even if I personally found this section a bit too short.

The focus of **Chapter 4** is on geographical variation, although O'Sullivan, Amador-Moreno & Barron emphasise that "[a] strict separation between geographical variation and other levels of variation cannot and should not be established" (p. 129). First, the concepts 'accent', 'dialect', 'variety', and 'idiolect' are introduced, then the authors briefly sketch central methods used in dialectology and variationist sociolinguistics and summarise four influential models of World Englishes (Kachru's Three Circles, Schneider's Dynamic Model, Buschfeld and Kautzsch's Extra- and Intra-territorial Forces model, and Mair's World System of Englishes). The main part of the chapter is dedicated to introducing corpora suitable for variationist sociolinguistic research, both in the field of World Englishes (Section 4.5) and for analyses of non-native Englishes (Section 4.6). Each corpus is briefly described with respect to its size and design, and for some corpora potential applications or disadvantages are pointed out. In addition, O'Sullivan, Amador-Moreno & Barron refer to the option of creating a specialised corpus for studying variation.

Whereas the previous chapters were concerned with corpus approaches to 'traditional' sociolinguistic

topics, **Chapter 5** addresses variation in language use online. Two main approaches are distinguished: first, the web can be used as a corpus itself, providing scholars with enormous amounts of data from varieties around the world. To access the web as a corpus, specialist software is required, one of which, WebCorp (RDUES 2022), is described in more detail. However, O'Sullivan, Amador-Moreno & Barron also address problems of the web-as-a-corpus approach, such as biases in search engines or inaccessible websites. In the second approach, the web is a data source for offline corpora, giving researchers more control over their data and the analysis process. To showcase how online corpora can be utilised to examine linguistic variation, studies using two web-based corpora, the Intelligent Web-based (iWeb) corpus and the corpus of Global Web-Based English (GloWbE) are presented. Still, O'Sullivan, Amador-Moreno & Barron conclude that, despite the obvious advantages of using huge online corpora for sociolinguistic research, the "dangers of depending on big data as the main source for building corpora and the value of carefully compiled corpora" (p. 172) must not be underestimated. In addition, it is not yet clear how exactly the interaction of various technological, social, contextual, and speaker-internal factors impacts online language as a separate genre with specific sub-genres (pp. 175–176). The chapter ends with an overview of central studies addressing the influence of context, social factors, and speaker agency on language use online and a section dedicated to ethical considerations in using online data.

The next two chapters have a historical focus. **Chapter 6** addresses corpus linguistic contributions to the analysis of language change. Specifically, O'Sullivan, Amador-Moreno & Barron illustrate how the creation of historical corpora has allowed researchers to access large data collections and how corpus analytical tools have "made it feasible for scholars to find patterns of variation and change by using these large amounts of data" (p. 194). Various historical corpora are described, including not only general but also specialised collections and corpora including data from non-British contexts. One section is dedicated to some of the challenges researchers might face when working with historical datasets, such as missing metadata, the underrepresentation of particular social groups, questions of authorship, or spelling variation leading to difficulties for automatic corpus searches. O'Sullivan, Amador-Moreno & Barron encourage readers to make use of complementary resources, for example, literary texts, dictionaries, linguistic atlases, or grammars and writing manuals. They argue that historical linguistic analysis must combine quantitative corpus analysis with qualitative analysis to provide "the contextual evidence that may sometimes be lost by simply interpreting structural patterns" (p. 211).

Chapter 7 is concerned with diachronic variation in the media. Three topics are covered: first O'Sullivan, Amador-Moreno & Barron address media influence on language change and provide an overview of corpus linguistic approaches to examine this issue (Section 7.1). Second, they discuss change in media language itself, for instance, towards increasing informality and greater diversity (Section 7.2). Again, suggestions for suitable corpora and concrete examples for studies on the topic are provided. Finally, Section 7.3 examines the relationship between language ideology and the media. Always referring to recent research in this area, O'Sullivan, Amador-Moreno & Barron illustrate how corpus linguistic tools, such as keywords or collocations, can help uncover both explicit and implicit language ideologies.

The conclusion in **Chapter 8** provides a concise summary of the previous chapters, addresses current issues in the field (such as insufficient metadata, ethical considerations, or challenges in working with big data), and touches upon future developments and desiderata (such as a more comprehensive annotation of corpora and the inclusion of video and audio recordings).

Overall, *Corpus linguistics for sociolinguistics* is an excellent source for lecturers to use in course preparation or for students and researchers wanting to improve their corpus linguistic skills. I was particularly impressed by the level of detail O'Sullivan, Amador-Moreno & Barron provide when introducing the various tools for sociolinguistic corpus analysis. The suggested activities allow readers to familiarise themselves with each tool, and the ensuing discussion questions can be used by lecturers to initiate discussions of both the respective results and the methods used.

Nevertheless, there are some aspects which should also be addressed. First, I was rather surprised to see little reference to socio-economic class and ethnic identity. Socio-economic class is addressed

briefly in the context of gender and swearing (pp. 60–63), but compared to other 'traditional' sociolinguistic variables, such as region, gender, or age, it receives considerably less attention. Ethnic identity is mentioned as one of the categories investigated in sociolinguistics, but except for some brief mentions in passing (e.g., Activity 4.10) the topic is not really taken up again in the rest of the book and does not even show up as an index term. Second, the book contains one factual mistake. The description of the spoken British National Corpus 2014 includes the statement "Recordings were conducted surreptitiously, without knowledge to participants" (p. 83). This is incorrect and probably due to a mix-up with the British National Corpus 1994, where informed consent was not yet required prior to the recordings (p. 84; cf. Love 2021: 746). Finally, the book contains a comparatively large amount of spelling mistakes, incomplete sentences, or similar editorial inconsistencies (e.g., inconsistent numbering of sections and line breaks mid-word), which could have been avoided by a bit more copy-editing.

These considerations aside, I wholeheartedly recommend *Corpus linguistics for sociolinguistics* for anybody interested in finding out more about how to harness corpus linguistics for sociolinguistic research. The book holds its promise of being an accessible guide into these fields and, in particular, of illustrating how they can fruitfully complement each other. This is achieved through the authors' lucid and engaging writing style as well as the book's hands-on approach and its successful inclusion of both classic and recent case studies.

References

Amador-Moreno, Carolina P. & Ana Maria Terrazas-Calero. 2022. Using corpus linguistics to explore literary speech representation: Non-standard language in fiction. In Anne O'Keeffe & Michael McCarthy (eds.), *The Routledge handbook of corpus linguistics,* 517–532. London: Routledge.

Biber, Douglas. 2012. Review article: *Sociolinguistics and corpus linguistics* (Paul Baker); *Corpus and sociolinguistics: Investigating age and gender in female talk* (Bróna Murphy). *Journal of Sociolinguistics* 16 (2). 301–307.

Love, Robin. 2021. Swearing in informal spoken English: 1990s–2010s. *Text & Talk* 41 (5–6). 739–762.

McIntyre, Dan & Brian Walker. 2019. *Corpus stylistics: Theory and practice*. Edinburgh: Edinburgh University Press.

RDUES = Research and Development Unit for English Studies. 2022. *WebCorp Linguist's Search Engine* (version 2). Birmingham City University. https://www.webcorp.org.uk/lse (Accessed 2025-12-29.)

Wu, Kan & Defeng Li. 2025. *Researching translator's functional style: A corpus-assisted approach*. Singapore: Springer.