

Review of Brookes, Gavin and Luke C. Collins. 2023. *Corpus Linguistics for Health Communication: A Guide for Research*. London: Routledge. ISBN: 978-1-003-09965-9
<https://doi.org/10.4324/9781003099659>

Ovidia Martínez Sánchez
University of Alicante / Spain

In this book, Gavin Brookes and Luke C. Collins introduce corpus linguistics for health communication from a research perspective. The monograph is included in the *Routledge Corpus Linguistics Guides* book series and significantly enriches the resources available to researchers in the field by 1) introducing the fields of health communication and corpus linguistics while critically evaluating recent studies in corpus-based health communication, 2) outlining the procedures involved in planning a corpus linguistic investigation of health communication, including corpus design and construction, tool selection, and implementation of analysis techniques, and 3) demonstrating the application and potential of corpus linguistic methods in the study of spoken, written, and digital health communication.

In terms of structure, the monograph consists of seven chapters, each of which is divided into different sections. Notably, all chapters follow a consistent organisational pattern, with an introductory section at the beginning and a summary section at the end, except for Chapter 7, designated as the conclusion. The chapter summaries are followed by a short list of suggested further reading, the authors' notes and the bibliography. What drew my attention was how the authors called this book an introductory textbook: "We have tried to write this book without assuming any prior knowledge of corpus linguistics or health communication" (p. 28). I was surprised by the premise that readers could easily understand the topics as they were introduced —as it happens alongside the book. I also liked the additional tasks for the readers related to the content presented in each chapter; thus, the book promotes interactive learning for researchers and students



alike. On top of that, the accessibility of the material is evident in the authors' careful attention to a modest and naturalistic stylistic approach.

The journey into the contents of the book begins with a thorough examination of Chapter 1, which introduces health communication and corpus linguistics and lays the groundwork for their significance and applications in the field. As emphasised by McCulloch *et al.* (2021: 28), “health communication is a multifaceted field of research, theory, and practise concerned with delivering health-related information to diverse populations.” When dealing with health communication and its definition, the authors also approach the nebulous term of ‘discourse’, drawing upon insights from various authors. We are also told that investigations of doctor-patient interactions have primarily dominated the study of health communication as a prototypical asymmetrical type of interaction (Linell 1990), traditionally characterised by different hierarchical positions on both sides, an aspect also explored in Chapter 4. Historically, the study of health communication primarily revolved around doctor-patient interactions, but a significant diversification emerged by the mid-1990s, encompassing interactions with other healthcare professionals such as nurses. Moving to corpus linguistics, in Section 1.3, they define a ‘corpus’ as a “machine-readable collection of authentic language use that has been sampled to represent a language or language variety” (p. 13). The chapter underscores the advantages of corpus linguistics in dissecting language and facilitating quantitative and qualitative insights into datasets. Notably, empirical corpus linguistics methods offer researchers in healthcare a data-driven approach to studying naturally occurring language use. This approach extends its benefits to diverse linguistic contexts, as evidenced by examples from the *RareDis Corpus* (Martínez-deMiguel *et al.* 2022), the *SetembroBR Corpus* (Ramos dos Santos *et al.* 2024), and the *EasyCall Corpus* (Turrisi *et al.* 2021). Section 1.5 candidly discusses limitations in using corpus linguistics approaches, shedding light on potential challenges for researchers. For instance, creating a comprehensive corpus is time-consuming, converting original texts can alter them, and automatically identifying some aspects like pragmatic features or metaphors is difficult. However, advancements in natural language processing provide promising solutions for those mentioned tasks. The chapter offers a concise summary clarifying key insights and prompting further exploration.

Chapter 2 starts with an introduction regarding its scope and outlines the difficulties of specialised corpus design and construction. First, Section 2.2 discusses

the considerations in designing a corpus, followed by practical aspects such as text collection (Section 2.3), cleaning (Section 2.4), and annotation processes (Section 2.5). The chapter explores the concepts of authenticity, text selection, representativeness, corpus size, and balance. It further discusses collecting spoken and written texts, detailing challenges and methodologies, including transcription complexities. Additionally, it also stresses the importance of cleaning and annotating the corpus for usability and reliability. Cleaning involves refining data by removing noise, standardising formats, correcting errors, handling duplicates, and filtering irrelevant content. Annotation adds metadata or linguistic tags like part-of-speech tags or syntactic structures to facilitate analysis. These processes ensure accuracy and usability for linguistic studies and natural language processing tasks. Tools such as *Sketch Engine* (Kilgariff *et al.* 2014) and *#LancsBox* (Brezina *et al.* 2015) are recommended for these tasks, both highly recognised and well-known in the field of corpus linguistics. Finally, the ethical considerations involved in corpus construction are addressed in Section 2.6, emphasising the importance of responsible research practices with texts of public consumption or informed consent when building corpora, such as spoken language corpora. In the summary section, the chapter offers a roadmap for researchers in health communication, advising the utilisation of existing corpora to save time and effort compared to creating new ones.

Chapter 3 ventures into the nature of corpus analysis, explaining the methodologies and tools required to glean meaningful insights from collected data, such as *AntConc* (Anthony 2022) and *WordSmith Tools* (Scott 2020). Beginning with an introduction, the chapter guides the landscape of software selection for corpus analysis. It then delves into various analytical techniques, including frequency, keyword analysis, collocation, cluster, and concordance analyses, each offering unique perspectives on the corpus data. Section 3.3. offers practical examples and illustrative explanations of the mentioned tasks with the *Daily Mail Dementia News Corpus* (Brookes 2023), and it demystifies the intricacies of each analytical approach. The chapter summarises the discussed methods, emphasising the indispensable role of human researchers in interpreting and positioning software tools' output, directly impacting results. Therefore, this chapter is readable for those who do not research health communication, but still want to learn and get introduced to the analysis of language in a corpus.

Chapter 4 focuses on the study of language in spoken health communication. Section 4.2 provides an in-depth analysis of data collection and analysis approaches and discusses the challenges posed by spoken language. It emphasises that the dialogic nature of spoken health communication is one of exchange, often involving the coming together of professional and patient expertise. It also introduces two case studies on the interactions between doctors and patients, from anorexia and register analysis. The approach to spoken health communication as a dialogue emphasises the sequence of turns. In response, researchers have often combined corpus linguistics with interactional approaches to analysis, such as conversation analysis. Section 4.3 then examines different forms of spoken health communication, from clinical interactions to research discourse and media representations. Section 4.4 analyses the characteristics of spoken health communication, including medical and interpersonal aspects. It highlights the importance of interactivity and the representation of different social actors in health and illness discourse. The chapter concludes with a summary to recapitulate the main points studied.

Chapter 5, the longest in the book, begins with an introduction that contextualises the main topic, namely, written forms of health communication and how these have been the subject of corpus analysis. Section 5.3 proceeds to explore the diverse forms of written health communication, including the genres of clinical documents, media texts, historical documents, and literary works. Here, I expected to find a greater variety of genres when dealing with written medical corpora. While genres like clinical guidelines, protocols, and medical reports are typically fundamental for examining communication practices across various medical fields, notably, they are neither used nor referenced in this chapter for corpus construction. The chapter analyses the distinctive characteristics of written health communication, examining the portrayal of health professionals and patients as well as the representation of illness, treatments, and solutions. It draws upon two case studies: one dealing with a longitudinal study of dementia metaphors in UK tabloids (Brookes 2023), and another revolving around examining lived experience through features of mind style (Demjén and Semino 2021). The authors' review of different types of written health communication in this chapter shows that representations of patients across texts are characterised by those affected by illness having limited agency. Furthermore, when people who are the subject of healthcare concerns are recorded in written texts, they are often depicted as lacking autonomy. The

chapter summary suggests incorporating patients' perspectives into corpus linguistics research on health texts to enhance the representation of first-person viewpoints in healthcare reports.

In the contemporary era, we have observed a remarkable proliferation of technological innovations, which has resulted in a profound transformation of how information is stored and accessed. The Internet and digital information are now frequently the first sources consulted by the public to obtain health information and for getting help with health-related questions or problems. The accessibility and quality of digital health tools may vary, resulting in disparities in health information access. Chapter 6 examines the methodological complexities of investigating medical language in the digital domain. It examines the various forms of digital health communication, including curated health information shared by professionals, interactive exchanges between individuals and health experts, and peer-to-peer health discussions facilitated by digital platforms. Section 6.3.2 introduces digital interactions with health professionals, mainly on social media platforms. The chapter provides a case study, by Hunt and Harvey (2015), about online discursive representations of eating disorders on the Internet through health queries. From my point of view, the case study is very engaging since it highlights that, in digital spaces, individuals can potentially express themselves in ways that are distinct from professional psychiatric contexts. As the dynamic between health and professional and patient shifts, the authors have observed patient perspectives that strive for autonomy and self-management regarding their health and medicalising discourses that appear to diminish individuals' sense of responsibility for their state of ill health. Moreover, a particular strength of digital spaces is their capacity for social connectedness and forming communities around shared health concerns, which can bring together members otherwise separated by time and space.

In this chapter, it has also been seen that their review of corpus studies of digital health communication is the discursive strategies through which contributors construct the persona of expert, administer advice, and provide readers with the content to extrapolate from their personal experiences. As an example, Coltman-Patel *et al.* (2022) carry out a key-word analysis of forum threads discussing vaccination, which directed them to focus on the use of insults, which were used as rhetorical devices. Their work attests to the value of extended and contextualised investigation of keywords that direct

us to the discursive aspects of deliberations around health concerns and the interpersonal dynamics of digital forums. Chapter 6 also shows how researchers have reported the potential advantages of the broader adoption of emojis in health documents, including capturing experiences of illness symptoms and health information (Lotfinejad *et al.* 2020). Section 6.4.3 deals with how natural language processing techniques may be used to process corpus and perform tasks related to sentiment analysis. In this section, however, I expected more examples of how to perform these sentiment analyses with some steps. The chapter ends with a summary, offering a comprehensive understanding of the different topics related to digital health communication.

Finally, chapter 7, which is the concluding chapter and the shortest in the book, reflects on the continued application of corpus linguistics methods to study health communication. The chapter considers how corpus linguistics can develop alongside advancements in healthcare to ensure that those using corpus linguistics methods can continue to make a meaningful contribution to the study of health communication and practice in the design and delivery of healthcare. This concise yet pivotal chapter embraces the positivist essence of both disciplines and assesses their vitality and prospects for the future. Section 7.2 considers the future trajectory of corpus-based health communication studies while acknowledging its transformative impact on the field.

Overall, *Corpus Linguistics for Health Communication: A Guide for Research* presents a thought-provoking perspective on health communication within corpus linguistics. The authors emphasise the importance of clarity and precision in research methodologies throughout the book, which are crucial considerations as the field expands quantitatively. This extraordinary volume offers a comprehensive set of case studies, tasks, website links, and an exhaustive list of further references, facilitating an understanding of the chapters' contents. It serves as an invaluable resource for those seeking to delve into the intersection of corpus linguistics and health communication. Including case studies throughout various sections provides concrete examples and practical applications of the discussed concepts. In addition, further reading sections at the end of each chapter provide additional material with brief notes from the authors on the importance of the recommended works. Thus, the reader is given a verified reference database of articles and books. Moreover, even for those unfamiliar with corpus methodological approaches to examining the language, this book can be an

excellent opportunity to read and learn the first steps to value this field. Suffice it to say that Gavin Brookes and Luke Collins have produced a comprehensive and accessible guide that will inform and inspire further research and exploration in the fields of corpus linguistics and health communication.

REFERENCES

- Anthony, Lawrence. 2022. *AntConc* (version 4.2.0). Tokyo: Waseda University. <https://www.laurenceanthony.net/software>
- Brezina, Vaclav, Tony McEnery and Stephen Wattam. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20/2: 139–173.
- Brookes, Gavin. 2023. Killer, thief or companion? A corpus-based study of dementia metaphors in UK tabloids. *Metaphor and Symbol* 38: 213–230.
- Coltman-Patel, Tara, William Dance, Zsófica Demjén, Derek Gatherer, Claire Hardaker and Elena Semino. 2022. Am I being unreasonable to vaccinate my kids against my ex's wishes? A corpus linguistic exploration of conflict in vaccination discussions on Mumsnet Talk's AIBU forum. *Discourse, Context & Media* 48: 100624. <https://doi.org/10.1016/j.dcm.2022.100624>
- Demjén, Zsófica and Elena Semino. 2021. Stylistics: Mind style in an autobiographical account of schizophrenia. In Gavin Brookes and Daniel Hunt eds. *Analysing Health Communication: Discourse Approaches*. Houndmills: Palgrave, 333–356.
- Hunt, Daniel and Kevin Harvey. 2015. Health communication and corpus linguistics: Using corpus tools to analyse eating disorder discourse online. In Paul Baker and Tony Mcenery eds. *Corpora and Discourse Studies: Integrating Discourse and Corpora*. Houndmills: Palgrave, 134–154.
- Kilgariff, Adam, Vit 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.
- Lotfinejad, Nasim, Reza Assadi, Mohammad Hassan Aelami and Didier Pittet. 2020. Emojis in public health and how they might be used for hand hygiene and infection prevention and control. *Antimicrobial Resistance and Infection Control* 9/27. <https://doi.org/10.1186/s13756-020-0692-2>
- Martínez-deMiguel, Claudia, Isabel Segura-Bedmar, Esteban Chacón-Solano and Sara Guerrero-Aspizua. 2022. The RareDis corpus: A corpus annotated with rare diseases, their signs and symptoms. *Journal of biomedical informatics* 125: 103961. <https://doi.org/10.1016/j.jbi.2021.103961>
- McCulloch, Seth, Grace M. Hildenbrand, Katie J. Schmitz and Evan K. Perrault. 2021. The state of health communication research: A content analysis of articles published in *Journal of Health Communication and Health Communication* (2010–2019). *Journal of Health Communication* 26/1: 28–38.
- Ramos dos Santos, Wesley, Rafael Lage de Oliveira and Ivandré Paraboni. 2024. SetembroBR: A social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation* 58: 273–300.
- Scott, Mike. 2020. *WordSmith Tools* (version 8). Stroud: Lexical Analysis Software.

Turrisi, Rosanna, Arianna Braccia, Marco Emanuele, Simone Giuliatti, Maura Pugliatti, Mariachiara Sensi, Luciano Fadiga and Leonardo Badino. 2021. EasyCall corpus: A dysarthric speech dataset. *Interspeech*: 41–45.

Reviewed by

Ovidia Martínez Sánchez

University of Alicante

Instituto de Lenguas Modernas Aplicadas

Edificio Institutos Universitarios II (Parque Científico)

03080 Alicante

Spain

E-mail: ovidia.martinez@ua.es