Cardiff University Press
Gwasg Prifysgol Caerdydd

JCaDS
JOURNAL OF CORPORA AND DISCOURSE STUDIES 2025, 9:56–63
E-ISSN 2515-0251

REVIEW: MCENERY T., & BREZINA, V. (2022). *FUNDAMENTAL PRINCIPLES OF CORPUS LINGUISTICS*. CAMBRIDGE UNIVERSITY PRESS.

SELENIA ANASTASI
UNIVERSITY OF GENOA

CITATION

Anastasi, S. (2025). Book Review: McEnery T., & Brezina, V. (2022). *Fundamental Principles of Corpus Linguistics*. Cambridge University Press . *Journal of Corpora and Discourse Studies*, *9*:56-63

CONTACT

Selenia Anastasi, Department of Modern Languages and Cultures, University of Genoa, via Balbi 5, Genoa, 16126, Italy. selenia.anastasi@edu.unige.it

DOI

10.18573/jcads.154

ISSUE DOI

10.18573/jcads.v9

ORCID

0000-0001-7696-7523

Manuscript accepted 2024-08-27

Book Review: McEnery, T., & Brezina, V. (2022). *Fundamental Principles of Corpus Linguistics*. Cambridge University Press.

Selenia Anastasi
University of Genoa

1. Introduction

What defines a method as scientific? Can Corpus Linguistics be considered a science? Can qualitative and quantitative approaches coexist fruitfully? These are the three main questions that *Fundamental principles of Corpus Linguistics* (McEnery & Brezina, 2022) aims to answer, proposing and articulating 48 principles in an incremental and iterative process, through which the authors equip the discipline with a necessary epistemological toolbox.

The text begins with a debate on the scientific nature of corpus linguistics, using as a pretext the provocations of Noam Chomsky, who questioned the foundation and methodological validity of the field (McEnery & Brezina, 2022, p. 6). In the course of the work, the authors not only respond in detail to Chomsky's criticisms and present several arguments in favor of the statistical study of language from corpora, but they also try to define the boundaries of the discipline.

Drawing on Karl Popper's philosophy of science (Popper, 1976, 2002), McEnery and Brezina (2022) propose a rigorous, non-dogmatic model of linguistic research which, starting from empirical data, experience and intersubjective observation (Principle 7), proceeds from the formulation of hypotheses to their falsification, based on principles of reproducibility and replicability of experiments (Principles 11 and 21). Necessary, to this end, is a balance between reliance on conventional methods (Principle 20) already attested in the literature — the only antidote to the problem of infinite regression in induction already addressed by Popper — and their continuous, necessary, problematising. During this brief review, I will proceed with a more detailed account of the 8 chapters, concluding with an exposition of some personal considerations that have emerged during my careful reading of the text.

The first two chapters of the text, *The First Sketch* (p. 3) and *What Is Science?* (p. 29), are devoted entirely to defining the position of the two authors in relation to scientific method; beginning with Chomsky's claims about the innateness of linguistic acquisition, they distinguish between scientific method based on critical realism and metaphysical approaches justified from the use of rhetoric. In this section, an initial and fundamental incompatibility emerges with Chomsky's view rooted in a perception of science inclined toward *scientiae rationalis*, that is, approaches that are based on logical reasoning and rational constructs. In contrast to Chomsky, in fact, McEnery and Brezina embrace a view of science as *scientia realis* (at the very fundaments of empiricism and *critical realism*), without neglecting the importance of rationality, but placing sensory experience and the

Anastasi (2025) Book Review: McEnery, T., & Brezina, V. (2022). Fundamental Principles of Corpus Linguistics. Cambridge University Press. DOI 10.18573/jcads.154

testability of a hypothesis at the heart of knowledge. This approach not only reflects a conception of science closer to that of the social sciences, as discussed further in the book, but also allows corpus linguistics to claim a legitimate place among scientific disciplines because of its ability to observe, measure and falsify hypotheses about language in use. All this is most elegantly summarized by the authors in Principle 6" (revised and refined throughout the text):

> Corpus linguistics, drawing on *scientia realis*, works, as a social science, in a way which is informed by concepts from science – it is the study of observable language on which experience may be tested in accordance with Principles 7 and 11. (pp. 256–257)

The third chapter turns to the development of useful operational guidelines for corpus linguistics, in accordance with what was discussed in Chapters 1 and 2 in relation to the scientific method. First, the authors implement an important distinction between inductive and *quasi-inductive* method, as it emerges from Popper's arguments. Here, a theory is scientific only when it can prove its verifiability and is thus validated by experimental corroboration. So, the initial deductive procedure leading to the development of the hypothesis should always be supplemented by bottom-up empirical verification of an inductive kind (i.e., from observations of individual cases to the theory being tested) (p. 107). This does not mean abandoning completely to rigid forms of positivism, depriving a theoretical system of openness to intuition in favor of static approaches anchored in tradition. On the contrary, good insights, when supported by observations of data, can lead to the development of original research hypotheses, which can in turn be corroborated through the integration of new tools and/or new data. Rather, what matters at this stage of confirmation of a theory or potential renewal of a system is the possibility of testing the theory so that previous assumptions can be reshaped to maximize the set of potential falsifiers while reducing the range of admissible statements, in accordance with Principle 28':

> Methods should be combined so as to: 1) maximise potential falsifiability; 2) maximise experimental falsifiers; 3) maximise the empirical content of the system under examination; and 4) minimise range. All of this is done in the pursuit of simplicity (p. 100).

Building on these assumptions, the chapter discusses the importance of consistency and non-contradiction of the claims allowed by any theoretical system. The set of acceptable claims must be stable — at least until the theory is revised and updated. As an example of this, the authors cite the long-standing debate over the unscientificity of topic modeling (Brookes & McEnery, 2019; Gillings & Hardie, 2023), which, lacking stability in generating outputs, '[is] metaphysical rather than scientific' (McEnery & Brezina, 2023, p. 83).

The chapter concludes by emphasizing the centrality of the researcher's intellectual honesty, a prerequisite without which any scientific research is rendered impossible. In fact, McEnery and Brezina argue that the researchers should first choose to persevere on the path of falsification rather than give in to the search for cases that confirm their worldview (and thus fall into the trap of *confirmation bias*).

The chapter provides a set of key concepts for understanding how, at the heart of corpus linguistics, lies a radically fallibilist view of science, framed as a position that 'recognises that our knowledge is not definitive because human beings are potentially fallible (hence the name) and our knowledge needs to be constantly revised' (p. 112). This argument is taken up and developed in more detail in Chapter 4, particularly with reference to the relationship between science and society and cultural reality, framing corpus linguistics in the broader context of the social sciences and, to some extent, the digital humanities.

Starting from Popper's idea that research can be divided into theoretical, historical and applied goals, the authors explain how corpus linguistics and the social sciences combine both scientific and interpretive approaches to understand the interconnectedness of physical and social aspects of reality. These goals can be achieved through a combination of naturalistic and anti-naturalistic perspectives. According to the authors, the situated dimension of linguistics, that is, included in a finite space and time, should not be seen as a limitation to its scientificity. It is in the inherently limited nature of corpus linguistics that lies the reliability of its method, in contrast to a metaphysical view of science centered on the search for 'divine Truths' that are untestable and therefore unfalsifiable. Indeed, according to Principle 35', 'A corpus, representing socially situated data, is an amalgam of social and physical interactions [and by virtue of this], our approach to analysing it should take this into account, recognizing the value of the theoretical, the historical and the applied' (p. 117). In other words, just as a model of an object cannot be an exact reproduction of the object it represents, so too corpus linguistics should not and cannot be concerned with constructing complete and exhaustive resources with which to arrive at universally acceptable truths about language. Rather, the task of corpus linguists is to provide results that, in their partiality, can represent small but important pieces in the mosaic of human understanding of society mediated by the language in use.

Since we do not have empirical resources that represent language use universally (a desirable goal only for those who adopt metaphysical approaches to reality), the authors explain the practices of corpus linguistics as a quasi-contact relationship with linguistic reality (critical realism).

In analysing reality through sensory experience, the researcher should take into account a certain degree of propensity. We can understand the concept of propensity in relation to that of probability. According to Popper, propensity implies that there are always certain conditions that can influence the probability of an event occurring. In other words, given conditions A-Z, these conditions have an intrinsic tendency B to produce certain results. Thus, instead of considering the probability of a word occurring in the

language as a whole, the researcher using corpora should be aware of the degree of propensity for a particular event to occur given specific circumstances A-Z. However, in corpus linguistics this is further complicated by the quasi-contact approach. While in a naturalistic method as well as in a controlled experiment (such as that of rolling a dice or a coin) we can access and measure propensity (considering, for instance, the shape of the dice and its weight), in the case of language in use, the set of forces that can influence observable events is not always identifiable.

The topic of 'language in context' is further discussed and expanded in Chapter 5, through the concept of 'everyday language use'. Here the authors again take their distance from the Chomskian universalist approach to language, explaining the importance of corpora for the understanding of real language in real contexts. This approach conceives communication as a means rather than an end, recognising the functionalist foundations of corpus linguistics through the principles of the Prague School and Halliday's functionalist linguistics.

By accepting corpus linguistics as a discipline associated with the social sciences, we recognise language production as a situated social action performed by subjects endowed with communicative purpose and intentionality. In this perspective, language production is always a performance and as such can be imperfect. In other words, a corpus is always a representation of the experience of subjects using language in ways that are appropriate to the context in which they find themselves. This principle applies not only when formulating a theory, but also when attempting to falsify its validity through what the authors call *in situ falsification*. In light of this, the authors formulate another important principle, Principle 41:

> The evidence for language is the production and reception of language by users and learners of language. The way in which it acts as evidence for them is a way in which it can act as evidence for linguists also (p. 161).

In other words, there is no privileged way of accessing linguistic knowledge other than through the experience of language itself. This principle should be considered particularly in relation to the development of ontologies and taxonomies for the annotation of linguistic data, which in turn are not to be seen as static photographs of an unambiguous reality, but as tools for scholars to frame and study linguistic phenomena from an angle of their interest.

Chapter 6 reiterates the centrality of the principles of repetition and replication for empirical studies. These two principles are central to corpus linguistics in that, in order to formulate new research hypotheses, the scholar should first critically explore the hypotheses covered by previous authors in an attempt to falsify them. The authors, therefore, proceed by defining the concept of repeatability as the ability to draw on data and tools already used in previous studies to confirm the results, within the limits granted by data retrieval, software availability and transparency of reported outcomes. To succeed in

these purposes, a good practice in empirical studies is to make the data on which the study is based available to the community. Sharing data sources allows for future verification and validation by other researchers working in the same area of investigation. This observation may seem trivial to those working in the so-called hard sciences, but in the field of linguistics and social sciences, problems relating to privacy and the management of personal data are often exacerbated by the scarcity of economic and human resources to cope with both the collection of data (which leads researchers to become jealous of their sources of analysis) and their maintenance in archives over long periods of time.

Regarding replicability, the authors defined it as 'a process in which our understanding based on prior research is confronted with some new empirical evidence consonant with a previous study' (McEnery & Brezina, 2022, p. 217). In corpus linguistics, replicability can be achieved through several expedients, the first of which involves replicating the research design on data that are collected in different time ranges. Leech's (2011) and Baker's (2017) experiments on the Brown corpus in relation to the use of modal verbs in British and American English are presented as examples of such a solution. Other approaches include integrating new data into a single corpus or applying the research design to multiple corpora. In general, it can be said that for the authors, the most effective way to falsify and replicate a study on corpora, and thus restrict the problem known in computer science as 'overfitting', is to validate the experiment on larger or comparable data samples. In this sense, the more applicable a theory is to different case studies, the more robust it is and can become part of the convention.

At this point in the book, the authors have provided the necessary pillars for the construction of the 48 basic principles of Corpus Linguistics. The last chapters (7 and 8) are therefore concerned with bringing the principles together by applying them to a case study to demonstrate how they work in practice, particularly in relation to the issue of replicability of experiments. This is accomplished by replicating Leech's study of modal verbs. Initially, the authors describe how they refined their hypotheses by subjecting them to increasingly rigorous testing, but without relying on statistical inferential tests, following the example of previous studies that, while recognizing the importance of statistical significance, do not rely on it to justify their claims. The text then raises questions about the use of significance tests and their epistemological value in corpus linguistics, questioning the appropriateness of being influenced by such tests and the real usefulness of these tests in the context of probability and the stability of observed statistical averages, proposing that they should be considered *methodological falsifications* that, while useful in practice, are not rigorously falsifiable because of the indirect nature of their connection to observations.

The authors conclude by advice readers to develop a rigorous methodological and epistemological approach to corpus linguistics, which should advance from review to review, and convention to convention, as new problems emerge based on the work done by the scientific community (as demonstrated by the authors during the development of the 48 principles).

I would like to conclude this review with some personal considerations in response to McEnery and Brezina's work. In particular, I would like to focus my attention on the notion of *socially situated knowledge* as discussed by the two authors, one of the cornerstones of the scientific approach to the study of corpora. In *Fundamental Principles of Corpus Linguistics*, the authors frequently refer to an epistemological conception not far from more recent feminist approaches to science. Indeed, the notion of *situated knowledge* (Haraway, 1991; Anderson, 1995) refers to, among other things, the fact that individuals can understand the same object in different ways from their distinct relations to the object itself. The researcher is also immersed in a network of relationships both with other scholars related to the same tradition and with the immediate object of their study, which in turn cannot be separated from the context of production. Both the knowing subject and the known subject are immersed in a spatiotemporal context that cannot be abstracted. So, although the authors do not explicitly refer to feminist theory, some continuity can be traced. This continuity becomes more clearly visible from the observation that the use of corpora, like any empirically grounded method, can provide only partial knowledge of the social context studied. Circumscribing the scope of our access to knowledge allows us to realize some of the assumptions assessed by feminist epistemology: 1. An *embodied* knowledge, aware of the spatiotemporal context in which the subjects-objects of knowledge find themselves *materially* acting and interacting; 2. An *experiential* knowledge, since through the use of corpora we recognize that we have access to a knowledge of the world that is always mediated — by language — by the physical and mental states that affect subjects; 3. We recognize that individuals involved in knowledge production may hold different beliefs about the same object, by virtue of different perspectives, theories, interests, and cultural background. Furthermore, we recognize that individuals involved in knowledge production may be in different epistemic relationships with both other scholars and other institutions that influence, or make possible, the transmission of such knowledge.

Ultimately, *Fundamental Principles of Corpus Linguistics* represents an essential reading not only in bringing together key issues from a methodological perspective in relation to corpora construction and analysis, but in providing appropriate critical tools for researchers who wish to approach the use of data without sacrificing the prolific coexistence of quantitative and qualitative approaches. Taken together, McEnery and Brezina's arguments make the fundamental principles of corpus linguistic natural allies for all those theoretical frameworks and academic traditions that reject the idea of 'knowledge or vision from nowhere' — a way of conceiving knowledge from a privileged point of view — in favour of more pluralistic approaches.

References

Anderson, E. (1995). Feminist Epistemology: An Interpretation and Defense. *Hypatia*, *10*, 50–84. https://doi.org/10.1111/j.1527-2001.1995.tb00737.x

Baker, P. (2017). A*merican and British English: Divided by a common language?* Cambridge: Cambridge University Press.

Brookes, G., & McEnery, T. (2019). The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies*, *21*(1), 3–21. https://doi.org/10.1177/146144561881403

Gillings, M., & Hardie, A. (2023). The interpretation of topic models for scholarly analysis: An evaluation and critique of current practice. *Digital Scholarship in the Humanities*, *38*(2), 530–543. https://doi.org/10.1093/llc/fqac075

Haraway, D. (1991). *Simians, Cyborgs, and Women*. New York: Routledge.

Leech, G. (2011). The modal verbs ARE declining: Reply to Neil Millar's 'Modal verbs in TIME: Frequency changes 1923–2006', *International Journal of Corpus Linguistics*, *16*(4), 547–564. https://doi.org/10.1075/ijcl.16.4.05lee

Popper, K. (1976). The logic of the social sciences. In Adorno, T. (Ed.), *The positivist dispute in German sociology* (pp. 87-104). London & Edinburgh: Heinemann.

Popper, K. (2002). *The logic of scientific discovery*. London & New York: Routledge.