
To date, only a few studies have been carried out on how to position the corpus linguistic approach within the study of language and within scientific approaches in general. Among the notable exceptions, McEnery and Brezina’s introduction (p. 1–2) mentions Leech (1992), Stubbs (2001) and Teubert (2005). There is thus a considerable gap to be filled by the present volume. The focus is not, however, to contrast the authors’ stance with that of previous linguistic studies, but instead to take renewed look at corpus linguistics through Karl Popper’s work on the philosophy of science, as this “provokes new ways of looking at old problems and practices” (p. 2).

Across the chapters, McEnery and Brezina formulate 48 principles of corpus linguistics, some of which are partly modified in the course of the discussion. These principles constitute the theoretical foundations of corpus linguistics – three of the central ones are given below:

- Principle 7: Corpus linguistics promotes and is based upon an intersubjectively observable approach to language in which results are repeatable and replicable. (p. 26)
- Principle 14: Corpus linguists are critical realists who use corpus data as a way to come into quasi-contact with language in use. (p. 14)
- Principle 23: Corpus data are as provisional as hypotheses and in principle may pass through a similar lifecycle. (p. 68)

Critical realism is here (p. 267) defined as “[a]n ontological position, which accepts that our contact with reality is not direct; […] one should be permissive of alternative perspectives and findings while pursuing one’s own perspectives and findings in the understanding that they may, or may not, prove in the long run to be sustainable.” Critical realism is a cornerstone of the authors’ take on (social) science. Throughout the first six chapters, McEnery and Brezina develop these 48 principles, highlighting the nature of knowledge attainable through corpora and how linguists can approach these limitations.

The first chapter starts by outlining how corpus linguistics fits into a general framework of science (cf. principle 7 above). Corpus linguistics tends to gravitate towards the empiricist corner (scientia realis) in its focus on observable language. Formal approaches to linguistics instead incline towards rationalism, highlighting logic and rational constructs (scientia rationalis). As expected, the authors distance themselves from Chomsky’s critical approach to corpus linguistics.

Chapter 2, “What is Science?”, takes on the question of how corpus linguistics adheres to the posited model of science. In science, and in linguistics in particular, we have to operate with provisional truths...
that can be falsified in the future. It is stressed, however, that science does not wait for the perfect data and method, so researchers have to work with what is available at a specific moment in time. The axioms (‘fundamental assumptions on which a theory is built’) of linguistics are axioms by convention, i.e., they rely on convention to operate. The conventions within linguistics are exemplified by such a basic concept as a ‘word’ being different across studies.

Chapter 3, entitled “How to do science”, deepens and widens the discussion of corpus linguistics, positioning it within the realm of social science rather than among the “hard” sciences. Linguistics largely relies on convention and language varies over time and space, hence the exploration of language has more in common with the social sciences. The authors concede that the data and measurements in corpus linguistic studies may turn out to be inadequate for the task at hand, but, as the authors argue, this also applies to any other branch of science. The chapter also problematizes the concept of falsification, cautioning against rejecting a hypothesis solely on the basis of a falsification. A crucial point is that falsifications can themselves be subject to falsifications.

The position of linguistics within social science is the topic of Chapter 4. According to Popper’s approach to science, we cannot know anything with absolute certainty, falsifiability is a core feature of science and the understandings we have are only provisional. This does not mean that anything can be true, but that some knowledge is more likely to be acceptable on conventional grounds.

In Chapter 5, “Everyday Linguistics: Form and Function”, the authors elaborate on the functionalist foundations of corpus linguistics. Language is seen as a purposeful means of expression, and linguistic structure is thought to be greatly shaped by the function of conveying meaning. In this context, “imperfect performance” from spoken language is discussed extensively. The nature of such material is particularly relevant as it has been used by formalist approaches to reject performance data. This part of the chapter largely consists of qualitative analyses of “in situ falsifications regarding beliefs of appropriacy and well-formedness of a linguistic feature or structure” (p. 163). Such falsifications may involve false starts corrected by the speaker or when a hearer overtly questions the well-formedness of an utterance. However, the authors indicate that a speaker may not always be committed to producing well-formed utterances, and a hearer may consider an utterance ill-formed without explicitly saying so. Reasonable assumptions about the purposefulness of communication can still be made, it is argued, but researchers have to be aware of and explicit about the fact that they are indeed making assumptions when making assessments. Such considerations are among the key differences between linguistics and physical sciences, the authors suggest.

Chapters 6 (“Repetition and Replication: Laying the Groundwork for an Empirical Study”) and 7 (“Replication: Carrying out an Empirical Study”) use Leech et al.’s (2010) data from the Brown family of corpora on English modal verbs as a point of departure on repetition and replication. McEnery and Brezina note that the repeatability of corpus studies is often limited due to 1) the availability of data, 2) the availability of software, and 3) the clarity of reporting. Illustrative examples are given of how different corpus versions – and different software – often yield different token frequencies from the same data files. Such issues indicate the importance of version management of corpora, for instance by saving old versions of corpora.

Chapter 7 starts with a lucid overview of the continuum of repetition of research to original study (p. 217): direct repetition of research (using the same data, tools and methods to test the accuracy of a previous study), repetition with variation (at least one variable changed, such as using a corpus with different annotation), direct replication (using new data, but the same methodology), broad replication (“[c]onfrontation of any aspect of previous research with new evidence”) and original study (testing a hypothesis for the first time). Most of the chapter focuses on the repetition of the study of modals. Particularly interesting in the comparisons with Leech et al.’s findings is that McEnery and Brezina compile their own versions of the 2006 data – two new each for American and British English – to test to what extent different corpus results are produced when using the identical sampling frame. In the end, Leech’s original results stand up “remarkably well” (p. 244) in the replication, but a few hypotheses are falsified in the scrutiny. For example, although modals generally decrease in the new datasets, the hypothesis positing that low-frequency modals are decreasing faster than high-frequency ones is
not borne out by the new data. Overall, it is notable that the different versions of the corpora produce similar, but far from identical results, so that individual modals typically vary between 10 and 20 per cent between the highest and the lowest frequencies.

The book ends with a concluding Chapter 8. This is followed by two appendices, one repeating the 48 principles in a condensed, accessible format (pp. 256–262) and one glossary (pp. 263–287) listing the key terminology of the work.

In conclusion, this volume is a major contribution in bringing together corpus linguistics with the philosophy of science. The book-length format allows extensive treatment of the complex issues at hand. Many corpus linguists will benefit from reading this work before embarking on their next project, whether it concerns a small-scale study or a major project involving corpus compilation. The position of corpus linguistics as a social science relying on observable language data puts specific demands on researchers within this paradigm. McEnery and Brezina make an excellent case outlining these, indicating the authors’ firm grasp of the subject. The book sets high (and sensible) standards for accountability in corpus linguistic studies: The corpora, tools and methods should be described in adequate detail in each study. This is a demand that may nevertheless be rather difficult to fulfil in the many highly condensed case studies published in journals and edited volumes, but this should certainly be an ideal to strive for. All in all, *Fundamental principles of corpus linguistics* will be of interest to any scholars doing corpus linguistic research.

## References


