Introduction

Corpus linguistics has, to date, largely been oriented towards the development of techniques for the analysis of texts and on the reporting of analyses undertaken with such techniques. By and large the approach taken has been what one might call 'common sense'. People have experimented with the construction of data sets and have developed ways of analysing texts using general good practice and by puzzling through issues in publications and debate. Overarching views of what the corpus approach constitutes in the study of language are rare – any works of our own to one side, probably the most notable examples of work in this vein are those of Leech (1992), Stubbs (2001a) and Teubert (2005).[1] Stubbs' work is closest to what this book does and, as it is discussed in Chapter 2, we will not mention it further here. Leech (1992) is also very close to the work presented here, as Leech was influenced by the work of Karl Popper, as we are. However, Leech, in the short space available to him, only focused on five features to characterise corpus linguistics – falsifiability, completeness, simplicity, strength and objectivity. Of those only falsifiability and simplicity are clearly drawn from the work of Popper, though Popper is directly referenced elsewhere in the piece and his shadow is clearly present throughout. Teubert, in contrast, provides, in a relatively short piece, a much fuller account of what he calls 'My version of corpus linguistics'. In many ways this book was inspired by that paper. Teubert lays out, in twenty-five theses, his view of what corpus linguistics constitutes. Teubert's view is relatively distinct from our own, though there are, inevitably, areas of overlap. However, the main difference starts with his first thesis, where he states that corpus linguistics 'is not concerned with the psychological aspects of language' (Teubert 2005: 2–3). We make no such claim in this book and the principles of corpus linguistics we lay out, grounded as they are in a form of quasi-contact with reality, are certainly permissive of making inferences about the mind. Similarly Teubert limits corpus linguistics to a study of language that is stored – again, we have a different vision of it, as a study of observed

[1] Though chapter one of Sampson (2017) also mines a similar vein of thought to that presented in this book.

language that enables comment on that which is not seen. Readers interested in a critical view, which we agree with, should see the comments on Teubert's theses in Gries (2012).

However, the purpose of this book is not to contrast our vision of corpus linguistics with those of other corpus linguists. The true aim of this book is to take the framework developed by one philosopher, Karl Popper, as it provokes new ways of looking at old problems and practices. Hence we will use his work as a prism through which we may view corpus linguistics, assessing the degree to which it can be viewed from that perspective and seeing what insights that view provides us. Along the way we will engage with other philosophers where it is useful to do so, but throughout it is Popper who is our guiding light. We use Popper's approach because, ultimately, it does lead to insights into the field. It also provides a framework and concepts that are helpful to corpus linguistics. With that said, for readers who wish to use other frameworks, or agree with other approaches to the subject, they may be interested in looking at this book to measure their views against ours, much as we were inspired by looking at the work of Teubert to lay out 'our' version of corpus linguistics. In doing so they will, we hope, be aided by the depth of treatment that we have provided for the reasoning behind each principle presented in this book, and our attempts to show those principles in action.

One final note is needed before we begin. When this book proposal was originally reviewed, the reviewers wondered whether, because of the nature of the enterprise we have set ourselves, we may be seen to be critical of the work of others. While that was never our intent, we have dealt with it in this book by trying, wherever possible, to highlight the work of ourselves or close colleagues when exemplifying points. So if any criticism is implied at any point, it will more likely apply to us and our colleagues than to other scholars.

CHAPTER ONE

The First Sketch

This book offers a journey in search of the theoretical foundations of corpus linguistics as an empirical science. It looks for the sources of knowledge about language and society that we can find in corpora (systematic samples of language) and how this knowledge can be evaluated and built upon. But why does this matter? Although we live in an era of unprecedented access to information via various online media (as of 2019, Wikipedia alone had over 50 million articles[2]) with scientists making use of ever larger data sets (big data), our knowledge base has often been shaken by false claims and systematic disinformation. In politics, terms such as 'post-truth' and 'fake news' have been widely circulated to further erode our trust in traditional epistemic authorities such as science (e.g. Ylä-Anttila 2018). The classical view that 'science should deliver the facts, and just the facts, needed for political decision-making, whereas liberal democracy should make decisions on the basis of these' (Kappel and Zahle 2017: 1) can no longer be seen as generally accepted. We therefore need to look at the role of science in the process of knowledge acquisition, critically reviewing the foundations of scientific practice and situating the endeavour of linguists and social scientists within this framework. In addition, this exploration is important because owing to its relatively short history dating back to the 1960s,[3] corpus linguistics has never made explicit the essential premises it is based upon, often being caught up in a discussion of whether it is principally a method or a theory of language (McEnery and Hardie 2011: 210ff.).

So what sort of foundations are we looking for? The foundations proposed in this book are arrived at by confronting corpus linguistic practice and its aspirations, on the one hand, with theoretical thinking about science (the philosophy of science), on the other. We start with a simple sketch based on a common-sense understanding of science,

[2] History of Wikipedia, https://en.wikipedia.org/wiki/History_of_Wikipedia.

[3] However, see McEnery and Hardie (2011: 37) for references to earlier sources of corpus linguistic techniques (concordances), which pre-date the invention of an electronic computer.

reality and truth, which will be further refined in later chapters of this book. The guide in the process of refining our initial ideas will be Popper's (e.g. 1972, 1975) work on the philosophy of science and its operations. We also draw on a number of examples from corpus linguistics and other disciplines (e.g. astronomy, biology, physics, psychology) and engage with comments and criticisms raised in relation to corpus linguistics by different scholars. Essentially, the foundations of corpus linguistics we are looking for are not completely solid, immovable rules, but a set of principles, considerations and guidelines that reflect the reality of corpus research. These foundations are similar, in a metaphorical sense, to the foundations of the Capitol Building in Utah in the United States of America (see Figure 1.1). These are foundations built not on rock but on shock-absorbing pillars able to withstand the impact of an earthquake. These foundations are strong enough to support an important building (corpus linguistics, in our metaphor), yet they are flexible enough to accommodate exogenous shocks and challenges from critique and research findings.

The method used in this exercise is one of heuristic engagement with corpus linguistics, philosophy of science and reality. Our claims are not

Figure 1.1 Seismic base isolation system under the State Capitol Building in Utah (source: https://commons.wikimedia.org/wiki/File:Base_isolators_under_the_Utah_State_Capitol.jpg)

absolute – in most cases, absolute claims lead to self-refutation. Instead, we seek to define and subsequently refine our understanding of the role of corpus linguistics as a discipline by building the contextualised groundwork for the application of corpus linguistics in the analyses of language and society. In this chapter, we sketch with broad strokes our initial position on corpus linguistics as science. We thus present what we believe is a common-sense approach to science. Based on the writings of others on the topic, and our own experience of it, we outline, informally, a description of what science appears to be. With this introduction to the idea of science, and an in-depth exploration of how it may apply to corpus linguistics established, we will then proceed to what will take up the bulk of this book: an exploration of the relationship between corpus linguistics and the ideas of Karl Popper regarding the scientific method. In appealing to this one framework we are, as we will show, drawing on a framework that many corpus linguists have touched upon, but not discussed extensively. We will also use that framework to move critically through a view of what constitutes science towards a view of what constitutes social science. This view will also encompass a variety of possible uses of corpus linguistics in digital humanities (e.g. Mahlberg and Wiegand 2020). After considering the nature of the data we interact with when we examine a corpus, we will conclude with a corpus-based study showing how the approach to social science we propose in this book – critical realism – applies to corpus linguistics through a consideration of the nature of statistically driven investigations of corpus data. While our focus will be on the framework established by Karl Popper (see the Key Thinkers box), we will also draw on the work of other philosophers, including Thomas Kuhn and Imre Lakatos, when discussing some of Popper's ideas and positioning ourselves relative to those.

Key Thinkers: Karl Raimund Popper (1902–1994)

Austrian-born British philosopher of science and social and political thinker. Born and educated in Vienna, died in London. Popper held academic positions at the University of Canterbury in New Zealand (1937–1946), the London School of Economics (1946–1949) and the University of London (1949–1969). His book *The Logic of Scientific Discovery*, originally published in German in 1934 and translated in English in 1959 with latest revision from 1972, is considered one of the most prominent works of the philosophy of science.[4]

[4] Note that in this book we make reference to versions of Popper's key works as reprinted in 2002 (Popper 2002a, 2002b, 2002c and 2002d) to ensure that readers will find it as easy as possible to refer back to the editions that we cite.

1.1 Linguistics and Science

The question of the relationship of linguistics to science has been very influential in the development of corpus linguistics. The move away from using observed language usage as evidence in linguistics was based, in part, on the claim that to use such observational data was not scientific, a claim still made by one of the linguists most closely associated with this claim, Noam Chomsky:

> Corpus linguistics doesn't mean anything. It's like saying suppose a physicist decides, suppose physics and chemistry decide that instead of relying on experiments, what they're going to do is take videotapes of things happening in the world and they'll collect huge videotapes of everything that's happening and from that maybe they'll come up with some generalizations or insights. Well, you know, sciences don't do this. (Chomsky as reported in Andor 2004: 97)

This quote gives a particular view of science. From it, we can gather that Chomsky believes that science should be based on experimental observation and that natural observations are not the stuff of science. We also may infer that while these claims are being made specifically with reference to physics and chemistry, what is said is meant to be true of the sciences in general as 'sciences don't do this'. As a result of views like these, corpus data, observed language use, was eschewed by mainstream linguistics for a long time and to an extent still is by some linguists. Hence, what constitutes science, and the permissibility or otherwise of corpus data, are issues which are very much linked, though rarely explored in depth. We should state clearly and without equivocation that we do not agree with Chomsky's view of what corpus linguistics constitutes nor do we accept his conception of what science is. That is not to say that we do not believe that there is space for experimentation and evidence in linguistics and in the sciences. There is and this book will argue for the need to bring research methods and data together to focus on research questions in order to explore them from a range of perspectives, insofar as this is possible and productive. Yet we do not accept that there is no place for natural observation in science; indeed, the observational sciences such as astronomy, epidemiology, geology and palaeontology routinely make systematic observations in order to make an assessment of theoretical claims. While they may also draw on other methods, systematic and structured observation is indispensable in these sciences. Science thus does not use only experiments under strict lab conditions to systematically collect data. In certain situations,

videotapes, which Chomsky condescendingly uses as a metaphor for measurements in real-world contexts, can provide indispensable and ecologically valid sources of data.[5] Linguistics, in our view, is the same. Corpus linguists do not aimlessly collect data as Chomsky suggested; however, we always need to be careful to critically evaluate the quality of corpora used in research. This leads us to our first principle.

Principle 1: Either when building a corpus or when using one, corpus linguists should use research questions in order to engage with their data in a structured and controlled way.

So, for instance, if you wish to look at language in informal spoken settings, you go to those settings and, using parameters that you believe may be relevant (different contexts, speakers of different ages, social classes etc.), you gather data and ensure that it encodes all of the contextual information you thought may be of interest. Afterwards, you use the structuring of the data to extract only the data that is relevant to your questions. So both when building a corpus and using it we are aware of the need to allow the data to respond to linguistically motivated, not aimless, enquiry.

Key Thinkers: Noam Chomsky (1928)

American linguist and social and political commentator. He has been a major influence on defining the programme of linguistics in non-empirical terms in the United States and internationally with his influential work *Syntactic Structures* (1957).

Chomsky is one of major critics of corpus linguistics, famously claiming that corpus-based 'description would be no more than a mere list' (Chomsky 1957: 159).

By the end of the book, we will have modified our claim and our categorisation of linguistics from a science to a social science. This does not change, nor is it caused by, our commitment to the permissibility of controlled observation of language, gathered from natural contexts of occurrence, as an important method in linguistics, science and the social sciences. Yet we will begin our journey to our categorisation of linguistics as social science by considering corpus linguistics and science not only

[5] Cf. Desagulier's discussion of this point and further debate with Chomsky, https://corpling.hypotheses.org/252.

because of how Chomsky has characterised it – as not being science – but also because of how some corpus linguists (e.g. Leech 1992) have characterised it – as science.

One of the claims that has been made to support the corpus approach to the study of language is that corpus linguistics is, or at least contains elements of, science. Leech (1992: 112) argues that corpus linguistics conforms to 'standards commonly applied to a theory in the scientific method' while de Beaugrande (1996: 533) argued that the main reason for using corpora was that it offered 'the first time in many years, a genuine opportunity to reorganise . . . doing language science'. McCarthy (2001: 125) makes the claim that corpus linguistics promises 'cutting edge change in terms of scientific techniques and methods"' while Stubbs (2001c: 154) draws parallels between corpus linguistics and the observational science of geology. Similarly, McEnery and Hardie (2011) appeal to the scientific method when arguing for the corpus approach to language. Yet the idea of the scientific method, as appealing as it sounds, is somewhat nebulous. This may sound surprising – surely scientists have a well worked out, non-contentious way of going about their studies? The truth of the matter is that the scientific method has been either heavily contested, as this chapter will show, or it is largely presented in the work of most scientists as a set of routinised procedures through which science can be conducted. A grand framework building up an overarching set of principles and practices by which all scientists work is a very nice idea, but, as with many ideas, has proved hard to achieve. For this reason, we build on simple principles such as common sense (see the next section), laying out the presuppositions behind those principles in the open and considering the role of inference and logic in building an empirical paradigm.

Let us begin by trying to draw out the main elements of what constitutes a scientific method. To slice through some of the indeterminacy, we will, as stated, begin with a common-sense approach to the study of science. That will allow us to sidestep some of the more esoteric points about the nature of 'truth' for example, to begin with, though we will return to that issue in this chapter and the issue will recur throughout the book. Having established a general framework for the scientific method, we will then explore the extent to which corpus linguistics does actually match it. The consideration of where the match may not be ideal then leads us on to our next chapter. To begin with, we will deal with a very simple conception of corpus linguistics, essentially based on looking up word forms and calculating frequency information. Later in the book,

when we move to problematise the engagement of corpus linguistics with the scientific method, we will consider a wider range of searches and techniques in corpus linguistics.

1.2 Realism and Common Sense

Let us begin with a common-sense definition of what science is. Science is the search for rational statements about reality. We appeal to common sense as the grounding for the position of realism taken in this chapter – human thought and individual physical actions and objects exist. Note that in appealing to common sense here we are setting aside, quite firmly, a series of positions that seek to challenge this view of science: scepticism, idealism and relativism. While we reject these as extreme positions, we accept elements of methodological scepticism and relativism (perspectivism) as healthy checks of our own thinking about reality. To deal briefly with each position, radical sceptics may question the very nature of reality itself, claiming that we can know nothing as nothing may exist, for example. However, our appeal to common sense here is designed to put what we view as flights of philosophical fancy to one side. We accept the existence of reality – as indeed radical sceptics tacitly do. We are reminded of a joke told to us by a colleague who studies Hinduism. A pupil is sat in a forest clearing with his guru who is explaining that all of reality is an illusion. At this point, an elephant emerges from the trees, charging at the pair. The guru stands and runs away, with his pupil following him. When the pair stop to catch their breath the pupil asks, 'Why did you run, the elephant was surely an illusion?' The guru replies, 'My running was an illusion too.' This neatly shows that sceptics seem to act as though reality exists – and quite rightly so too in our view. Sceptical positions can be easy to lampoon, as we, and others, have – Augustine of Hippo reports the difficulty of one prominent sceptic, Carneades, faced with the apparently difficult question of whether he was a man or a bug.[6] Yet while such positions can be viewed as risible, it took someone of Augustine's stature to be able to mount such an obvious attack because sceptical ideas, especially when linked to elites such as distinguished

[6] See King (1995: 71). Carneades made two amendments to his reasoning to deal with such an attack. The first was to say that his scepticism only applied to theoretical and philosophical matters. Second, he adopted the valuable notion of acting on the basis of what is probable rather than true. See the Introduction to *Against the Academicians and the Teacher* (King 1995: x–xi).

philosophers or religious teachers, have great appeal. So-called privileged cynicism[7] is one of the standard fallacies of reasoning, where highly sceptical views are accepted simply by virtue of being associated with a perceived elite. While critical thinking – the application of methodological scepticism – is a useful guide to querying hypotheses and testing them, so-called ontological scepticism defies common sense all too often, hence we set it aside here.

We also distance ourselves from idealism – the position that only the mind exists and objects and actions are imagined by the mind. This was the proposition of thinkers such as George Berkeley who, when proposing the possibility that a stone existed in the mind only, was famously confronted by Samuel Johnson who kicked the stone and said 'I refute it thus'. While Johnson was actually missing Berkeley's point, we would refute the argument on the lines of common sense – it beggars belief that this is the case. Our everyday practice also points to the reality of the world around us: it is a common-sense assumption that underlines our interaction with other people (who are outside our mind as clearly shown by the possibility of misunderstanding them) and objects around us (which we treat as real).

Finally, we set aside relativism, that truth is for the individual only hence no shared and objective truth is possible. Again, we appeal to common sense – in our everyday practice we assume convergence on some basic truths, which guide us as rational actors in the world. We assume, for instance, that if we drop an object, it will fall down to the ground. We also make a distinction between facts and opinions, the former being undisputable while the latter are open to debate and interpretation. Note that at this stage, we intentionally do not problematise this distinction, although examples of sensory deceptions and measurement errors can blur this line. Later (in Principle 14), we further develop this point within the framework of critical realism, which acknowledges that we have quasi-contact with reality. As with scepticism, we allow the possibility of methodological relativism (or perspectivism), which recognises the complexity of reality and our imperfect grasp of the truth. This allows for multiple perspectives, or interpretations, to compete in a rational debate when searching for the truth.

Dismissing such fine philosophical positions on the basis of common sense may seem to be rather high handed. However, a study of the scientific literature reveals that scientists are much more concerned

[7] See Pirie (2015) for a thorough review of this and other fallacies.