18

What can a corpus tell us about discourse?

Gerlinde Mautner

1 Corpora and discourse: mapping the terrain

Corpus linguistics (CL) and discourse studies (DS) evolved roughly at the same time, in the 1980s and 1990s, in what was originally a parallel development. There were few, if any, connections between them. Most corpus linguists were involved in lexicography and grammar, aiming to discover and quantify patterns in ever-larger corpora. Most discourse analysts, by contrast, focused on small samples of text which were suitable for close reading, "thick" description, and qualitative analysis sensitive to social context. 'For some considerable time,' Partington (2004a: 11) points out, 'the dichotomy was virtually complete: corpus linguists were generally unaware that their quantitative techniques could have much to say about discourse, while discourse analysts rarely saw reason to venture forth very far from their qualitative ivory tower'. In those early days, then, the question posed in the chapter title would rarely have been asked, let alone answered.

Now, in the third decade of the twenty-first century, the situation looks very different. The boundaries between the two approaches to language have become more porous, and their 'synergic use' (Partington and Marchi 2015: 219) is now quite common. In fact, a new, combined field has emerged, generally referred to as *corpus-assisted discourse studies*, or *CADS*. The field has consolidated (Mautner 2019), and there is a growing community of linguists who specifically identify as practising this strand of research.

The development sketched earlier needs to be put into perspective, however. The narrative of CL and DS initially leading separate lives never entirely reflected reality, at least not as starkly as Partington's (2004a) original quote may lead us to believe. To a present-day researcher using corpora for DS, it is both interesting and humbling to see how many decades before CADS, as such, was conceived, linguists were very much aware of the social component of corpus-based work. One of the earliest and probably best-known statements to this effect was Firth's plea, as early as the mid-1930s, for 'research into the detailed contextual distribution of sociologically important words, what one might call *focal* or *pivotal* words' (Firth 1935 [1957]: 10; original italics). In this

 DOI: 10.4324/9780367076399-18

quotation alone, the cornerstones of what was to become CADS are all there: *distribution* and *words* point to CL's strength in identifying lexical patterns, while *sociologically important* hints at the readiness of DS to be driven by relevance in the "real" world. And, of course, Firth's *focal or pivotal words* foreshadow keyword analysis, now a staple of CADS work.

Yet it was not until the 1990s that these ideas resurfaced and were developed further. Arguably the first paper to demonstrate how corpora could be studied from a social perspective was Leech and Fallon (1992). Tellingly, the title of their paper – echoed by the title of the present chapter – was "Computer corpora – What do they tell us about culture?" The idea of applying CL in this way gradually gained momentum. Stubbs and Gerbig (1993) investigated the use of verbs in geography textbooks; Caldas-Coulthard (1993) explored gender representation; and Hardt-Mautner (1995) used a study of European discourse in the British press to make a case for combining CL and discourse analysis, specifically its critical variety. It became apparent even then that the combination was not simply useful and interesting but had profound implications for all stages of the research process, and indeed for the nature of linguistic inquiry itself. Reflections in this vein are included, among others, in Stubbs (1996), Hunston (2002: 109–23), O'Halloran and Coffin (2004) and Mautner (2009, 2016).

Discourse: an elusive foundational concept

It speaks volumes that a term as central and widely used as *discourse* should still require clarification at all. As Thornbury (2010: 270) observed in the first edition of this handbook, '[t]he term *discourse* is both slippery and baggy: slippery because it eludes neat definition, and baggy because it embraces a wide range of linguistic and social phenomena'. There is general agreement that language plays a key role in it, but not everyone wishes to stop there. For Blommaert, for example (2005: 3), discourse 'comprises all forms of meaningful semiotic human activity seen in connection with social, cultural, and historical patterns and development of use'.

The picture becomes even more complex if we look across to the social sciences in general and individual disciplines, such as organisation studies. The so-called "linguistic turn" (Rorty 1992) has made *discourse* a term that is both *en vogue* and vague. It has such a wide range of meanings that Alvesson and Kärreman (2000: 1127) conclude: 'we cannot help sometimes feeling that the word discourse is used to cover up muddled thinking or postponed decisions on vital analytical matter'. And they continue: 'Discourse sometimes comes close to standing for everything, and thus nothing'.

In linguistics, however, the situation seems to be more straightforward. There, too, *discourse* can have fuzzy conceptual boundaries, but it is generally taken as read that at its core the term has two interconnected meanings: (i) stretches of language larger than single sentences and (ii) language embedded in social contexts and used to perform social functions. And whatever specific "brand" of discourse analysis a researcher subscribes to, they are unlikely to identify as working on "discourse" unless their research designs include both of these elements, "larger units" and "social context", in some form or other. Identifying specifically as a *critical* discourse analyst generally involves an interest in 'the semiotic dimensions of power, identity politics and political-economic or cultural change in society' (Wodak 2011: 38). With a brief as broad and complex as this, it is actually hard to see how we could do without "discourse". After all, as long as its meaning is not stretched too much, the concept has substantial explanatory power.

Whether explicitly of a critical persuasion or not, researchers investigating discourse typically ask research questions such as: In the texts that I am analysing, who are the main social actors, how are they described and what activities are they shown to be engaged in? How do the speakers/writers of these texts signal their own stance as well as social and professional identities? Can the patterns that emerge from the corpus be mapped on to and explained by the political, organisational and legal constraints under which the texts were produced? The actual questions asked will differ from project to project, but what unites all such undertakings is that the research journey begins and ends with text in context. Ideally, from the initial framing of the research design to the final interpretation of the data, the driving force is curiosity about how language and social reality are related. And at all stages of the journey, corpus analytic procedures can make a significant contribution.

Another attraction of CL is that it can be paired with any "brand" of discourse analysis, thus supporting a wide range of perspectives and methodologies. Some approaches may be a better fit than others, but it is hard to envisage a discourse analytical research design that cannot benefit at all from CL input. CL is flexible and adaptable. The only theoretical assumption the discourse analyst has to "buy into" to be comfortable with CL is that micro-linguistic choices and the macro-phenomenon of discourse are in a mutual relationship.

The intensity with which synergies are exploited may vary, of course. A study may be conceived as corpus assisted from the very outset, or it may merely dip into CL methods in order to confirm or refute a diagnosis reached by other means. Whatever the approach, there is a strong rationale for integrating corpus linguistics and discourse analysis, and this will be addressed in the next section.

2 The rationale for using corpus linguistic methods in discourse analysis

The two main methodological questions that first led to the emergence of corpus-assisted discourse studies are still the point of departure for most projects in this area today. How can we analyse discourse without relying too much on the researcher's intuition? And given that the traditional toolkit of discourse analysis is qualitative, how can we get a handle on our data when the corpus is too large to permit qualitative analysis? The answer to the first question is by working with a large and representative corpus of texts, by employing discovery procedures that are transparent and replicable and by relating one's findings to other corpora and thus putting them into perspective. The answer to the second is by employing CL techniques. It is not the only possible answer – with computer-based content analysis being a strong contender – but CL is the one that concerns us here. It has been widely shown to produce robust and insightful research exploring a range of questions at the discourse and society interface.

From the outset, the guiding principle of this partnership was to combine 'the quantitative rigour of corpus linguistics with the social perspective of qualitative approaches to discourse analysis' (Marchi and Taylor 2018: 4). The idea is not to replace qualitative approaches, but to put them on a sounder empirical footing. This, of course, addresses criticism of DS generally, and of critical discourse analysis in particular, that it is impressionistic and subjective and "cherry-picks" texts that are convenient rather than typical (cf. Widdowson 1995, 2004).

At the same time, there are heuristic benefits (i.e. related to how we arrive at new knowledge). 'A main aim of CADS', Subtirelu and Baker explain (2018: 108), 'is to

discover non-obvious meaning and identify questions and puzzles that would otherwise have not been considered' (Subtirelu and Baker 2018: 108). A corpus-based perspective on discourse thus not only enriches existing qualitative methods but also encourages us to put different questions to our data.

3 Using corpus linguistic tools to explore themes, genres and perspectives

There is a large and growing body of research that studies discourse with the help of corpus linguistic tools. The social domains and issues covered include politics (e.g. Partington 2011; Jeffries and Walker 2012; and Chapter 42, this volume), the law (e.g. Fanego and Rodríguez-Puente 2019; Chapter 41, this volume), business (e.g. Lischinsky 2011; Mullany 2013; Mautner and Learmonth 2020), gender and sexuality (Baker 2013; Zottola 2018), health communication (Hunt and Harvey 2015; Chapter 43, this volume) and online communication and social media (e.g. Lutzky and Kehoe 2016; Collins 2019; Chapter 46, this volume). All of these domains and issues could be (and are) also studied without recourse to CL. Yet CL tools provide added value in the sense described in the previous section: making larger datasets manageable, strengthening empirical claims and preventing the biased selection of texts for analysis.

So, what exactly are these tools and what can they contribute to the study of discourse? The most commonly used ones are frequency lists of both individual lexical items and n-grams (i.e. recurring word clusters of two words or more); keywords (i.e. words which occur statistically more or less frequently in one corpus than another); collocations (i.e. the words that commonly co-occur with a particular item); and concordances, which show search words in their surrounding co-text (see Chapters 9, 10 and 14, this volume). Concordances enable the analyst to examine the social domains that lexical items refer to, as well their "semantic prosody", that is, whether their evaluative orientation is positive or negative (Louw 1993; Hunston 2004: 157; Partington 2004b).

In what follows, I will explain how using each of these tools can tell us something about discourse. By way of examples, I will refer to two ongoing projects of mine, a study of UK Supreme Court judgements, on the one hand, and of articles in a management journal, on the other. The *UK Supreme Court* (UKSC) *Corpus* comprises all the judgements between 2009 and 2018 that have at least one dissenting opinion in them (129 in total, amounting to 2.87 million words). The corpus of management writing consists of 3,547 articles and book reviews from the *Administrative Science Quarterly* (ASQ), amounting to 15,885,378 million words. To enable diachronic comparisons across this latter corpus, it is divided into six subcorpora of roughly equal size, with material from the 1950s and 1960s put together in one subcorpus, and the other six decades each forming their own. For this illustrative analysis, both the UKSC and ASQ corpora were uploaded to Sketch Engine, a fee-based, internet-based concordancing programme (www.sketchengine.eu). The two corpora represent very different genres, produced in different institutional settings and enacting different social functions. What unites them, however, is that they both contain highly specialised, argumentative prose written and read mainly by professionals. The two projects were not conceived with a comparison between them in mind, nor is the purpose of this section to describe them in full. Even so, looking at selected features in two different corpora will help illustrate and highlight the potential of CL to enrich discourse analytic research.

Frequency

At first glance, simple wordlists may seem too pedestrian to be of much analytical value to the discourse analyst. However, they can be useful in several ways, especially if we focus on lexical rather than grammatical items (i.e. on nouns, verbs, adjectives, etc. rather than articles and prepositions). Even the raw frequencies of lexical items give us a rough idea of what the salient themes in our corpus are and how they are framed. In some respects the quantitative evidence may only confirm what we already know, but in others the results will be less obvious. It will hardly come as a surprise, for example, that the ten most frequent content words in the UKSC corpus are all legal terms: *court, case, section*, *para* (short for *paragraph*), *act, Lord, law, right, state,* and *appeal*. What would have been harder to predict, however, is that half of these legal terms describe other texts (*case, section, para, act, article*). In the ASQ corpus, three meta-discursive nouns make it to the top ten most frequent content words, namely *study*, *research* and *model* (the status of *work*, also in the top ten, is ambiguous, as it may refer both to research and to paid employment). Thus, even before we have done any in-depth analysis at all, even a small portion of the frequency list tells us something about legal and academic discourse that could be worth following up, namely that the former is heavily intertextual and the latter frequently uses meta-discursive elements. Frequency alone cannot tell us how these elements are used, whether they are self-referential (talking about the texts themselves) or intertextual (talking about other texts). But at the very least, we have made a first step towards profiling our corpus.

N-grams can also be a good starting point (see Chapter 15, this volume). Sketch Engine offers users the choice to view 2- to 6-grams, all of which are potentially interesting in their own right. Generally speaking, the shorter the n-gram, the less distinctive of a particular corpus it is likely to be. For example, a typical 2-gram frequently found across any corpus is *of the*. The longer the n-gram, on the other hand, the more likely it is to be corpus-specific. Two of the high-frequency 6-grams from the ASQ and UKSC corpora are, respectively, *from the point of view of,* a stock phrase from academic discourse, and *the European Convention on Human Rights,* a key document around which many Supreme Court cases revolve. Depending on the specific research question, then, both very generic and very distinctive n-grams may be useful. For illustrative purposes, it seemed best to go for the middle ground. Hence I computed 4-grams for both the UKSC and ASQ corpora. As always, some of the results are underwhelming – predictable and effectively dead-ends because one cannot really take them any further. For example, in the USKC corpus, the most frequent 4-grams are all names of institutions which are bound to occur in legal proceedings taken to the level of the highest court, such as *the Court of Appeal* and *the Secretary of State*. Yet on both lists, many n-grams appear to be much more general. Phrases such as *in the course of* or *are more likely to* sound like prefabricated chunks that could occur anywhere. Except that many of them do not, in fact, occur anywhere, but are characteristic of one corpus but not another. To take our example, with the 4-grams in the ASQ and UKSC corpora, it seemed reasonable to expect a fair amount of overlap because both corpora represent formal, argumentative prose. Yet among the 40 most frequent 4-grams (to take a random cut-off point), only 8 occur in both corpora. Thus, the use of n-grams can be shown to reflect genre conventions – which, in turn, are shaped by speech communities and their discursive practices.

One of these genre conventions is that the court judgements – also referred to, tellingly, as *opinions* – are framed discursively as the judges' personal views and as links in

an argumentative chain that includes lower-instance courts as well as other Supreme Court judges. That much we could also have ascertained by qualitative means through a close reading of individual judgements. We would probably also have noticed the following features: hedging devices and politeness markers (e.g. *I regret that I am unable to agree with his conclusion; I respectfully disagree; I respectfully agree but would add that …*); passages where judges engage directly with other judges' views (e.g. *I do not share his confidence about this*); meta-linguistic markers guiding the reader through the argument (e.g. *before considering these issues I should mention some other matters by way of background*); and rhetorical questions that give the texts an almost dialogic feel (e.g. *Can it be said that his decision would be immune from challenge? Surely not.*).

The added value of CL techniques is that they can help us substantiate such impressions and turn them into generalisable statements – and do both of these things very efficiently. Frequencies never give us the whole story, but they often provide interesting clues. To follow up on our impression that Supreme Court judgements sound personal, let us return to our n-grams. Among the 4-grams which occur in the top 40 of the UKSC corpus (but not the ASQ), two stand out in particular: *it seems to me* and *I agree with Lord*. Given the highly formal nature of judgements, the first person pronouns come as a surprise (at least to the analyst with a Continental-European background, where judgements give the opinion of the court and not of individual judges). The normalised, or relative, frequency of *it seems to me* is 131.5 occurrences per million words. As the next step, we may want to cast our net wider and search for *it seems to me* and *I agree with* in the *British National Corpus* (BNC), a large corpus suitable as a so-called "reference corpus", which reflects general usage (see Chapters 4, 10 and 39, this volume). In the BNC, it turns out, the two phrases are both considerably more frequent in the spoken subcorpus of the BNC than in the written one. (*It seems to me* occurs 28 times per million words in the BNC's spoken component, but only 5 times per million words in the written one; for *I agree with* the figures are 17 per million and 4 per million, respectively.) If we concluded from this evidence alone that the judgements use "spoken language", that would obviously be misguided and a case of rather simplistic "overinterpretation" (O'Halloran and Coffin 2004). Yet these frequencies do add a piece to the jigsaw puzzle, so that we can gradually build up the bigger picture and explain in more precise terms how the style, or "tone", that is typical of a genre actually comes about.

Thus, information about frequency is not only an interesting entry point into the data but can also yield substantive results that are relevant in their own right. To give another example, in Mautner and Learmonth (2020) we focus on lexical items that represent social actors in the ASQ. Comparing the normalised frequencies of selected social actor labels decade by decade, we found that across the time span covered, between 1956 and 2018, some social actor labels became more frequent (e.g. *CEO, entrepreneur* and *team member*), while others became less frequent or disappeared altogether (e.g. *administrator, bureaucrat, foreman, subordinate* and *supervisor*). These frequencies match what we know about more general socio-political and institutional trends in the last few decades (Learmonth and Morrell 2019). Broadly speaking, in management writing, bureaucratic elites and hierarchical relationships appear to have fallen out of favour, whereas labels that reflect managerialism, equality and neo-liberal rhetoric are now more popular. It ought to be stressed, however, that this narrative cannot be "read off" the corpus directly, but results from linking quantitative evidence with qualitative background knowledge. For the necessary interpretative act, the quality criterion is not "truth", but plausibility.

Keywords

Although normalised frequencies are a very useful measure, the comparison between a corpus and a reference corpus is best served by the computation of keywords (see Chapters 9 and 10, this volume). The software will flag up lexical items as "key" if the difference in frequency is statistically significant. Keywords are a robust measure that tells us what is distinctive, in lexical terms, about one corpus as opposed to another. Word choice is central to discursive construction, which, in turn, is central to social life. Thus, keyword analysis also allows us a glimpse of the similarities and differences between the socio-cultural contexts in which the texts in the corpus were produced.

Using a diachronic corpus, divided into subcorpora according to time periods, we can apply keyword analysis to explore how lexical choices have changed over time. To take an example from the *ASQ* study (Mautner and Learmonth 2020), we may want to compare the latest subcorpus, comprising articles published between 2010 and 2018, with the oldest subcorpus, dating from the 1950s and 1960s. The latter would serve as the so-called "reference corpus". Some of the words that appear as keywords are entirely predictable because their referents simply did not exist 60 years ago, such as *software* and *online*. Others refer to social roles and phenomena that certainly existed back then but were not conceptualised and written about in the way that they are now. Three items in that category are among the 50 items with the highest "keyness" scores: *CEO, gender* and *lesbian*. We can also "flip" the comparison, using the 2010s subcorpus as the reference corpus. The items that emerge as keywords for the older subcorpus include *foreman*, *bureaucracy* and *morale*, which ties in with our earlier diagnosis that the older subcorpus includes more terms associated with traditional business administration.

Like all other CL tools (and indeed empirical tests of any kind), keyword and frequency analysis can produce both interesting and boring results. Telling which is which, however, is not always entirely straightforward. "Interesting" is often associated with unexpected findings, and "boring" with predictable, "so what" kind of results (Baker and McEnery 2015: 9). Yet the latter, too, can be of value. On the one hand, they can substantiate what we may have suspected all along but were not able to back up empirically, and on the other they may lead to new observations that are not in fact that obvious.

There is no question that some disappointing results remain just that. But it is not uncommon for data to yield up their secrets much later, after a great deal more thinking and digging. The thinking part is about trying to work out why the corpus does not appear to be "behaving itself". The answer may be linguistic, connected to the lexico-grammatical properties of the items in question. Or it may lie outside the corpus itself, at the interface of discourse and institutional settings. The "digging" part typically involves studying collocations and concordances, which we will address now.

Collocations

The concept of collocation takes us back to Firth (1957: 11), who defined collocations as 'the mere word accompaniment, the other word-material in which [words] are most commonly or most characteristically embedded'. That "word-material" provides us with a wealth of information about discourse. Perhaps most significantly, it allows us to pinpoint how a speaker or writer expresses their stance towards certain people and events, how the roles of social actors are framed discursively and how they construct their own identities.

Studying collocations can also help us make sense of quantitative findings that appear not to add up – either because they are counterintuitive or at odds with information gleaned from other sources. In the ASQ study, we found that the frequencies for *manager* and *leader* had remained more or less constant during the period investigated, even though all our background reading around the subject suggested very strongly that *leader* should have become more common over time. If the difference did not lie in their frequency of occurrence, then it had to lie in *how* they were used. We therefore examined the collocates of the lemmata *manager* and *leader* in each of the subcorpora and within a span of three words to the left and right. (*Lemma* refers to 'a base form of a word together with its inflected forms' [Collins 2019: 197].) We noticed that, overall, *manager* and *leader* share few collocates (Mautner and Learmonth 2020: 284). The ten most frequent collocates for *manager* in the earliest subcorpus are labels indicating areas of responsibility (e.g. *departmental, district* and *sales*), as well as positions in a hierarchy (e.g. *assistant*). From the 1970s onwards, further hierarchical labels come in, including *top*, *middle* and *senior*. The collocates of *leader*, on the other hand, are originally terms associated with the public domain (e.g. *community, political* and *legislative*). From the 1980s onwards, *team* appears as a frequent collocate, and in the most recent corpus, leading up to 2018, *leaders* are typically labelled as *corporate*. Thus, while the frequencies of *manager* and *leader* have not changed significantly over time, their collocational profiles have, reflecting different and changing perspectives of these roles.

Concordances

Up to this point in the research process, the analytical output has been quantitative and focused on individual words or short phrases. We have seen that such information can help us identify dominant discursive constructions in large corpora. However, if we want to see language come to life, we need to go beyond individual words and examine propositions, that is, descriptions and evaluations of situations, people and activities. This is where concordances come in. By showing us search words surrounded by their co-text, they take us much closer to the original texts than do computational procedures. Naturally, in using concordances, we are still at some distance from the originals, but we are as close as we can get if we want to work with a corpus too large for manual analysis. In our *leader* and *manager* example, concordancing these two focal terms in the most recent subcorpus opens a window on contemporary organisational life (as seen through the lens of management scholarship, given the nature of the corpus). We learn from the concordances that: ***leaders*** *are expected to set moral examples for their followers; employed men in traditional marriages tend to (…) perceive organizations with female* ***leaders*** *as relatively unattractive; the prevalence of narcissistic* ***leaders*** *in American corporations may be a direct product of the society's prevalent individualistic culture;* ***managers*** *are typically cut off from the process of generating novel ideas;* ***managers*** *are likely to develop feelings of resentment toward the CEO; and once* ***managers*** *become desperate, they may act aggressively to remedy their problem quickly.*

Often the concordance lines contain words that catch our interest specifically in light of what we know about the socio-political background of our corpus. In our example, these could be *society*, *culture*, *female*, *effectiveness*, *resentment* and *frustrated*, to name just a few. For these expressions we could produce separate concordances, examine their frequency in the same or other subcorpora or search for them in reference corpora such as the *Corpus of Contemporary American English* (*COCA*, https://www.english-corpora.org/coca/).

Quite literally, one thing will lead to another in a process that should be equally focused and playful.

At the stage in the research process where concordances are analysed, CL edges towards the qualitative end of the methodological spectrum, while also maintaining its link with a quantitative view, which one can easily go back to if required. What Thornbury (2010: 282) calls 'cyclical alternation between counting and interpreting' is indeed at the heart of corpus-based discourse research. Whichever metaphor one uses to describe this process – whether it is "oscillating" (Mautner 2007: 66) or "shunting" (Partington and Marchi 2015: 231) – the idea is the same, describing a back-and-forth movement between quantitative and qualitative views of the data and triangulating one with the other. Ultimately, this approach is meant to bridge the divide between these two modes of analysis, or perhaps even dissolve it.

4 Epistemological issues

Epistemology boils down to a question that sounds deceptively simple: How do we know what we know? Earlier in the chapter, we argued that bringing in CL puts DS on a sounder empirical footing because it enables us to work with large datasets and delays the point at which interpretation sets in. Yet like all methods – including purely quantitative ones – the application of CL in discourse research raises a range of epistemological issues. I believe that these do not undermine the approach, but unless they are confronted, they may compromise the quality of the research.

The main challenges are these (Mautner 2016: 174–6). First, although quantitative corpus evidence can be seductive, it is never self-explanatory. As soon as you start explaining the evidence, which research is meant to do, subjectivity is inevitably back on the scene. The software lists frequencies for us and computes the strength of collocational attraction (see Chapters 9, 10, 15 and 13, this volume, on statistical measures of collocational strength), but it tells us nothing about what any of this means in a wider social context and why these linguistic choices were made in the first place.

Second, in interpreting findings, the temptation to jump to conclusions can prove almost irresistible. Yet it is not enough for the researchers themselves to claim connections between linguistic evidence and social meaning; they need to succeed in making others see these connections as well. Hence my appeal earlier in this chapter to keep a watchful eye on plausibility.

Third, even if the results are entirely convincing, we should always be careful not to overgeneralise from them. Every corpus, no matter how large and representative, is selective and an artefact in its own right. In making claims about the discourse concerned, or even language in general, it is always best to err on the side of caution.

Fourth, amid frequency tables and concordances, it is easy to lose sight of the complete texts that went into the corpus and the social context they were lifted from. Egbert and Schnur (2018: 172) are right to warn that 'discourse analysts must be wise in their use of corpus data and methods to ensure that the text retains its rightful status as the fundamental discourse unit'.

Fifth, statements about frequency must be comparative (frequent in relation to what?). Care must also be taken not to describe results in pseudo-quantitative terms (using vague descriptors such as *most*, *many* or *few* without backing them up with actual figures). In the qualitative phase of the analysis, involving concordances or complete texts, the attempt to quantify findings is often misplaced.

Sixth, when assessing the utility of the method, a degree of scepticism and humility is always a useful corrective. Discourse analyses that use CL are certainly different from those that do not, but they are not inherently superior. Nor are they necessarily the best fit for all research questions.

Seventh and finally, CL methods are not a "wonder drug" that will cure every ill that a research design may be suffering from, such as bias built into the corpus, through skewed sampling, for instance.

Overall, we have to be aware that, in spite of its benefits, 'the corpus approach in itself does not remove bias' (Baker 2018: 270). Even its "number-crunching" elements invariably involve decisions on the part of the analyst, such as which items to focus on, which corpora to compare and which cut-off points to use when it comes to deciding what is an important result. These concerns should not dampen our enthusiasm for employing CL tools, but should encourage critical reflexivity and healthy scepticism. Most importantly, we should not allow ourselves to be lured into thinking that employing corpus methods makes research completely "objective". Rather, it ought to be viewed as 'a means of achieving greater precision, richness as well as awareness' (Marchi and Taylor 2018: 6).

5 What a corpus cannot tell us about discourse

Methodological reflection and critique is as much about what a method cannot do as about what it can do. The potential of computer assistance in discourse studies is considerable, but it should not be oversold. For example, some "dusty corners" and "blind spots" remain (Marchi and Taylor 2018: 9). While CL has traditionally focused on differences between corpora and the presence of items, a strong case has been made to concentrate more on similarity (Taylor 2018) and absence (Partington 2014; Duguid and Partington 2018). Likewise, a preoccupation with high frequencies may not be suitable when the discourses to be explored are marginalised – those of or about minorities, for example – and for that reason leave few traces in texts (Motschenbacher 2018: 167).

Another area that is notoriously under-researched is multimodality (see Chapter 7, this volume). In building purely text-based corpora, we strip texts of pictures, sound and the haptic dimension, all of which may be essential for how texts are understood. Similarly, if spoken language is reduced to transcripts of what was said, important clues conveyed by facial expression and gesture will be lost, as will be information on how meaning unfolds in interaction and how it is co-constructed dynamically by the participants. The same applies to the study of metaphor. It has been shown to be amenable to corpus approaches (e.g. Deignan 2005; Deignan and Semino 2010), but there are bound to be areas that these cannot reach, such as the role that metaphors play in establishing cohesion across distant parts of texts.

In fact, quite generally, corpus techniques are not the "go-to" method when we are interested in what happens below the surface of the text (Thornbury 2010: 275). At least for standard CL procedures, we need lexical items that we can search for in the first place. It is only natural that this simple and incontrovertible fact should privilege some perspectives on discourse but hamper or even prevent others. Thus, to return to the question posed in the chapter's title, a corpus can tell us a lot about discourse, but not everything. But then, which method can? To unpack the intricate connections between

text and context, we will always need a variety of methods as well as creative, joined-up thinking to integrate them.

Further reading

Baker, P. (2006) *Using Corpora in Discourse Analysis*, London: Continuum. (This book remains a classic and a good starting point for those embarking on their first corpus-assisted discourse analysis. It combines hands-on tips with theoretical reflection and methodological critique.)

Collins, L. C. (2019) *Corpus Linguistics for Online Communication: A Guide for Research*, London: Routledge. (Although the book's focus is on online communication, it is also instructive for those doing corpus-assisted research on discourse in other areas. It also contains a glossary and tasks, with commentaries in an appendix.)

Taylor, C. and Marchi, A. (eds) (2018) *Corpus Approaches to Discourse: A Critical Review*, London: Routledge. (This edited volume is a thought-provoking account that experienced researchers are likely to find particularly useful. In three parts, it examines hitherto overlooked areas, triangulation and questions of research design.)

References

Alvesson, M. and Kärreman, D. (2000) 'Varieties of Discourse: On the Study of Organizations Through Discourse Analysis', *Human Relations* 53(9): 1125–49.

Baker, P. (2006) *Using Corpora in Discourse Analysis*, London: Continuum.

Baker, P. (2013) 'From Gay Language to Normative Discourse: A Diachronic Corpus Analysis of Lavender Linguistics Conference Abstracts 1994-2012', *Journal of Language and Sexuality* 2(2): 179–205.

Baker, P. (2018) 'Language, Sexuality and Corpus Linguistics. Concerns and Future Directions', *Journal of Language and Sexuality* 7(2): 263–79.

Baker, P. and McEnery, T. (2015) 'Introduction', in P. Baker and T. McEnery (eds) *Corpora and Discourse Studies. Integrating Discourse and Corpora*, Basingstoke: Palgrave MacMillan, pp. 1–19.

Blommaert, J. (2005) *Discourse. A Critical Introduction*, Cambridge: Cambridge University Press.

Caldas-Coulthard, C. R. (1993) 'From Discourse Analysis to Critical Discourse Analysis: The Differential Re-Presentation of Women and Men Speaking in Written News', in J. M. Sinclair, M. Hoey and G. Fox (eds) *Techniques of Description: Spoken and Written Discourse*, London: Routledge, pp. 196–208.

Collins, L. C. (2019) *Corpus Linguistics for Online Communication: A Guide for Research*, London: Routledge.

Deignan, A. (2005) *Metaphor and Corpus Linguistics*, Amsterdam and Philadelphia: Benjamins.

Deignan, A. and Semino, E. (2010) 'Corpus Techniques for Metaphor Analysis', in L. Cameron and R. Maslen (eds) *Metaphor Analysis: Research Practice in Applied Linguistics, Social Sciences and the Humanities*, London: Equinox, pp. 161–79.

Duguid, A. and Partington, A. (2018) 'Absence: You Don't Know What You're Missing. Or Do You?' in C. Taylor and A. Marchi (eds) *Corpus Approaches to Discourse. A Critical Review*, London: Routledge, pp. 38–59.

Egbert, J. and Schnur, E. (2018) 'The Text in Corpus and Discourse Analysis. Missing the Trees for the Forest', in C. Taylor and A. Marchi (eds) *Corpus Approaches to Discourse: A Critical Review,* London: Routledge, pp. 159–73.

Fanego, T. and Rodríguez-Puente, P. (eds) (2019) *Corpus-Based Research on Variation in English Legal Discourse*, Amsterdam: Benjamins.

Firth, J. R. (1935 [1957]) 'A Synopsis of Linguistic Theory, 1930-1955', in *Studies in Linguistic Analysis*, Oxford: Blackwell, pp. 1–32.

Firth, J. R. (1957) 'A Synopsis of Linguistic Theory 1930-55', in *Studies in Linguistic Analysis, Philosophical Society*, Oxford, reprinted in Palmer, F. (ed) (1968) Selected Papers of J.R. Firth, London: Longman, pp. 168–205.
Hardt-Mautner, G. (1995) '"Only Connect": Critical Discourse Analysis and Corpus Linguistics', UCREL Technical Paper 6, Lancaster: University of Lancaster. Available at http://ucrel.lancs.ac.uk/papers/techpaper/vol6.pdf.
Hunston, S. (2002) *Corpora in Applied Linguistics*, Cambridge: Cambridge University Press.
Hunston, S. (2004) 'Counting the Uncountable: Problems of Identifying Evaluation in a Text and in a Corpus', in A. Partington, J. Morley and L. Haarman (eds) *Corpora and Discourse*, Bern: Peter Lang, pp. 157–88.
Hunt, D. and Harvey, K. (2015) 'Health Communication and Corpus Linguistics: Using Corpus Tools to Analyse Eating Disorder Discourse Online', in P. Baker and T. McEnery (eds) *Corpora and Discourse Studies. Integrating Discourse and Corpora*, Basingstoke: Palgrave Macmillan, pp. 134–54.
Jeffries, L. and Walker, B. (2012) 'Key Words in the Press: A Critical Corpus-Driven Analysis of Ideology in the Blair Years (1998-2007)', *English Text Construction* 5(2): 208–29.
Learmonth, M. and Morrell, K. (2019) *Critical Perspectives on Leadership: The Language of Corporate Power*, London: Routledge.
Leech, G. and Fallon, R. (1992) 'Computer Corpora - What Do They Tell Us About Culture?', *ICAME Journal* 16: 29–50.
Lischinsky, A. (2011) 'In Times of Crisis: A Corpus Approach to the Construction of the Global Financial Crisis in Annual Reports', *Critical Discourse Studies* 8(3): 153–68.
Louw, B. (1993) 'Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies', in M. Baker, G. Francis and E. Tognini-Bonelli (eds) *Text and Technology. In Honour of John Sinclair*. Amsterdam: Benjamins, pp. 157–76.
Lutzky, U. and Kehoe, A. (2016) 'Your Blog is (the) Shit: A Corpus Linguistic Approach to the Identification of Swearing in Computer Mediated Communication', *International Journal of Corpus Linguistics* 21(2): 165–91.
Marchi, A. and Taylor, C. (2018) 'Introduction: Partiality and Reflexivity', in C. Taylor and A. Marchi (eds) *Corpus Approaches to Discourse. A Critical Review*, London: Routledge, pp. 1–15.
Mautner, G. (2007) 'Mining Large Corpora for Social Information: The Case of *Elderly*', *Language in Society* 36(1): 51–72.
Mautner, G. (2009) 'Corpora and Critical Discourse Analysis', in P. Baker (ed.) *Contemporary Corpus Linguistics*, London: Continuum, pp. 32–46.
Mautner, G. (2016) 'Checks and Balances: How Corpus Linguistics can Contribute to CDA', in R. Wodak and M. Meyer (eds) *Methods of Critical Discourse Analysis*, London: Sage, pp. 154–79.
Mautner, G. (2019) 'A Research Note on Corpora and Discourse: Points to Ponder in Research Design', *Journal of Corpora and Discourse Studies* 2: 1–13.
Mautner, G. and Learmonth, M. (2020) 'From *Administrator* to *CEO*: Exploring Changing Representations of Hierarchy and Prestige in a Diachronic Corpus of Academic Management Writing', *Discourse and Communication* 14(3): 273–93.
Motschenbacher, H. (2018) 'Corpus Linguistics in Language and Sexuality Studies. Taking Stock and Looking Ahead', *Journal of Language and Sexuality* 7(2): 145–74.
Mullany, L. (2013) 'Corpus Analysis of Language in the Workplace', in C. Chapelle (ed.) *The Encyclopedia of Applied Linguistics*, Chichester: Wiley-Blackwell, pp. 1–9.
O'Halloran and Coffin, C. (2004) 'Checking Overinterpreation and Underinterpretation: Help from Corpora in Critical Linguistics' in A. Hewings, C. Coffin and K. O'Halloran (eds) *Applying English Grammar*, London: Arnold, pp. 275–97.
Partington A. (2004a) 'Corpora and Discourse, a Most Congruous Beast', in A. Partington, J. Morley and L. Haarman (eds) *Corpora and Discourse*, Bern: Peter Lang, pp. 9–18.
Partington (2004b) 'Utterly Content in Each Other's Company: Semantic Prosody and Semantic Preference', *International Journal of Corpus Linguistics* 9(1): 131–56.
Partington, A. (2011) '"Double-speak" at the White House: A Corpus-Assisted Study of Bisociation in Conversational Laughter-Talk', *Humor: International Journal of Humor Research* 24(4): 371–98.
Partington, A. (2014) 'Mind the Gaps. The Role of Corpus Linguistics in Researching Absences', *International Journal of Corpus Linguistics* 19(1): 118–46.

Partington, A. and Marchi, A. (2015) 'Using Corpora in Discourse Analysis', in D. Biber and R. Reppen (eds) *The Cambridge Handbook of English Corpus Linguistics*, Cambridge: Cambridge University Press, pp. 216–34.

Rorty, R. (ed.) (1992) *The Linguistic Turn. Essays in Philosophical Method*, Chicago: Chicago University Press.

Stubbs, M. and Gerbig, A. (1993) 'Human and Inhuman Geography: On the Computer-Assisted Analysis of Long Texts', in M. Hoey (ed.) *Data, Description, Discourse. Papers on the English Language in Honour of John McH Sinclair on his Sixtieth Birthday*, London: Harper Collins, pp. 64–85.

Stubbs, M. (1996) *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*, Oxford: Blackwell.

Subtirelu, N. C. and Baker, P. (2018) 'Corpus-Based Approaches', in J. Flowerdew and J. E. Richardson (eds) *The Routledge Handbook of Critical Discourse Studies*, London: Routledge, pp. 106–19.

Taylor, C. (2018) 'Similarity', in C. Taylor and A. Marchi (eds) *Corpus Approaches to Discourse: A Critical Review*, London: Routledge, pp. 19–37.

Taylor, C. and Marchi, A. (eds) (2018) *Corpus Approaches to Discourse: A Critical Review*, London: Routledge.

Thornbury, S. (2010) 'What Can a Corpus Tell Us About Discourse?', in A. O'Keeffe and M. J. McCarthy (eds) *The Routledge Handbook of Corpus Linguistics*, London: Routledge, pp. 270–87.

Widdowson, H. (1995) 'Discourse Analysis: A Critical View', *Language and Literature* 4(3): 157–72.

Widdowson, H. (2004) *Text, Context, Pretext: Critical Issues in Critical Discourse Analysis*, Oxford: Blackwell.

Wodak, R. (2011) 'Critical Discourse Analysis', in K. Hyland and B. Paltridge (eds) *Continuum Companion to Discourse Analysis*, London and New York: Continuum, pp. 38–53.

Zottola, A. (2018) 'Transgender Identity Labels in the British Press: A Corpus-Based Discourse Analysis', *Journal of Language and Sexuality* 7(2): 237–62.