Using Corpora in Discourse Analysis

Paul Baker

2006

continuum
LONDON • NEW YORK

1 Introduction

This book is about a set of techniques of analysing language for a particular purpose. Or more precisely, it is about using *corpora* (large bodies of naturally occurring language data stored on computers) and corpus processes (computational procedures which manipulate this data in various ways) in order to uncover linguistic patterns which can enable us to make sense of the ways that language is used in the construction of *discourses* (or ways of constructing reality).

It therefore involves the pairing of two areas related to linguistics (corpora and discourse) which have not had a great deal to do with each other for reasons I will try to explain later in this chapter. This book is mainly written for 'linguists who use corpora' (Partington 2003: 257), rather than explicitly for corpus linguists, although hopefully corpus linguists may find something of use in it too.

This chapter serves as an overview for the rest of the book. A problem with writing a book that involves bridge-building between two different disciplines, is in the assumptions that have to be made regarding a fairly disparate target audience. Some people may know a lot about discourse analysis but not a great deal about corpus linguistics. For others the opposite may be the case. For others still, both areas might be equally opaque. So, for the sake of completeness and inclusiveness, I will try to cover as much ground as possible and hope that readers bear with me or can skim through the parts that they are already familiar with. I will begin by giving a quick description of corpus linguistics, followed by one of discourse.

Corpus Linguistics

Corpus linguistics is 'the study of language based on examples of real life language use' (McEnery & Wilson, 1996: 1). However, unlike purely qualitative approaches to research, corpus linguistics utilizes bodies of electronically encoded text, implementing a more quantitative methodology, for example by using frequency information

about occurrences of particular linguistic phenomena. As Biber (1998: 4) points out, corpus-based research actually depends on both quantitative *and* qualitative techniques: 'Association patterns represent quantitative relations, measuring the extent to which features and variants are associated with contextual factors. However functional (qualitative) interpretation is also an essential step in any corpus-based analysis.'

Corpora are generally large (consisting of thousands or even millions of words), representative samples of a particular type of naturally occurring language, so they can therefore be used as a standard reference with which claims about language can be measured. The fact that they are encoded electronically means that complex calculations can be carried out on large amounts of text, revealing linguistic patterns and frequency information that would otherwise take days or months to uncover by hand, and may run counter to intuition.

Electronic corpora are often annotated with additional linguistic information, the most common being part of speech information (for example, whether something is a noun or a verb), which allows large-scale grammatical analyses to be carried out. Other types of information can be encoded within corpora – for example, in spoken corpora (containing transcripts of dialogue) attributes such as sex, age, socio-economic group and region can be encoded for each participant. This would allow language comparisons to be made about different types of speakers. For example, Rayson *et al* (1997) have shown that speakers from economically advantaged groups use adverbs like *actually* and *really* more than those from less advantaged groups, who are more likely to use words like *say*, *said* and *saying*, numbers and taboo words.

Corpus-based or equivalent methods have been used from as early as the nineteenth century. The diary studies of infant language acquisition (Taine 1877; Preyer 1889), or Käding's (1897) frequency distribution of sequences of letters in an 11 million word corpus of German focused on collections of large, naturally occurring language use (in the absence of computers, the data was painstakingly analysed by hand). However, up until the 1970s, only a small number of studies utilized corpus-based approaches. Quirk's (1960) *Survey of English Usage* began in 1961, as did Brown and Kucera's work on the Brown corpus of American English. It was not until the advent of widely available personal computers in the 1980s that corpus linguistics as a methodology became popular. Johansson (1991) shows that the number of such studies doubled for every five year period between 1976–1991.

Corpus linguistics has since been employed in a number of areas of linguistic enquiry, including dictionary creation (Clear *et al* 1996),

as an aid to interpretation of literary texts (Louw 1997), forensic linguistics (Wools and Coulthard 1998), language description (Sinclair 1999), language variation studies (Biber 1988) and language teaching materials (Johns 1997). The aim of this book, however, is to investigate how corpus linguistics can enable the analysis of discourses. With that said, the term *discourse* has numerous interpretations, so the following section explains what I mean when I use it.

Discourse

The term *discourse* is problematic, as it is used in social and linguistic research in a number of inter-related yet different ways. In traditional linguistics it is defined as either 'language above the sentence or above the clause' (Stubbs 1983: 1), or 'language in use' (Brown and Yule 1983). We can talk about the discourse structure of particular texts. For example, a recipe will usually begin with the name of the meal to be prepared, then give a list of ingredients, then describe the means of preparation. There may be variants to this, but on the whole we are usually able to recognize the discourse structure of a text like a recipe fairly easily. We would expect certain lexical items or grammatical structures to appear at particular places (for example, numbers and measurements would appear near the beginning of the text, in the list of ingredients, e.g. '4 15ml spoons of olive oil', whereas imperative sentences would appear in the latter half, e.g. 'Slice each potato lengthwise.'). The term *discourse* is also sometimes applied to different types of language use or topics, for example, we can talk about political discourse (Chilton 2004), colonial discourse (Williams and Chrisman 1993), media discourse (Fairclough 1995) and environmental discourse (Hajer 1997). A number of researchers have used corpora to examine discourse styles of people who are learners of English. Ringbom (1998) found a high frequency of lexis that had a high generality (words like *people* and *things*) in a corpus of writings produced by learners of English when compared to a similar corpus of native speakers. Ringbom suggests that this results in learner English having a vague style. Similarly, Lorenz (1998) found that learners modify adjectives frequently, giving their discourse a sense of overstatement 'The sea was very clean', whereas Flowerdew (2000) showed that learner discourse contained an under-use of hedging devices (words like *perhaps* and *possibly*), making their writing appear overly direct. So this is a conceptualization of discourse which is linked to genre, style or text type. And throughout this book we will be examining a range of different discourses: tourist discourse in Chapter 3, news reporting

discourse in Chapters 4 and 7, and political discourse in Chapter 6. However, discourse can also be defined as 'practices which systematically form the objects of which they speak' (Foucault 1972: 49) and it is this meaning of discourse which I intend to focus on in this book (although in practice it is difficult to consider this meaning without taking into account the other meanings as well).

In order to expand upon Foucault's definition, discourse is a 'system of statements which constructs an object' (Parker 1992: 5) or 'language-in-action' (Blommaert 2005: 2). It is further categorized by Burr (1995: 48) as 'a set of meanings, metaphors, representations, images, stories, statements and so on that in some way together produce a particular version of events ... Surrounding any one object, event, person etc., there may be a variety of different discourses, each with a different story to tell about the world, a different way of representing it to the world.' Because of Foucault's notion of practices, discourse therefore becomes a countable noun: *discourses* (Cameron 2001: 15). So around any given object or concept there are likely to be multiple ways of constructing it, reflecting the fact that humans are diverse creatures; we tend to perceive aspects of the world in different ways, depending on a range of factors. In addition, discourses allow for people to be internally inconsistent; they help to explain why people contradict themselves, change position or appear to have ambiguous or conflicting views on the same subject (Potter and Wetherell 1987). We can view cases like this in terms of people holding competing discourses. Therefore, discourses are not valid descriptions of people's 'beliefs' or 'opinions' and they cannot be taken as representing an inner, essential aspect of identity such as personality or attitude. Instead they are connected to practices and structures that are lived out in society from day to day. Discourses can therefore be difficult to pin down or describe – they are constantly changing, interacting with each other, breaking off and merging. As Sunderland (2004) points out, there is no 'dictionary of discourses'. In addition, any act of naming or defining a discourse is going to be an interpretative one. Where I see a discourse, you may see a different discourse, or no discourse. It is difficult, if not impossible, to step outside discourse. Therefore our labelling of something as a discourse is going to be based upon the discourses that we already (often unconsciously) live with. As Foucault (1972: 146) notes, 'it is not possible for us to describe our own archive, since it is from within these rules that we speak.'

To give a couple of examples, Holloway's (1981, 1984) work on heterosexual relations produced what Sunderland (2004: 58) refers to as a 'male sexual drive' discourse, one which constructs male sexuality as a biological drive – men are seen as having a basic need for sex

which they cannot ignore and must be satisfied. Such a discourse could be used in law courts to ensure that male rapists receive lighter sentences. Similarly, Sunderland (2004: 55) identifies a discourse of *compulsory heterosexuality*, based on Rich's critical essay 'Compulsory Heterosexuality and Lesbian Existence' (1980). This discourse would involve practices which involve overlooking the existence of gay and lesbian people by assuming that everyone is heterosexual. Traces of this discourse could be found in a wide range of language contexts – for example, at a (traditional) wedding when relatives tell single people 'It'll be your turn next!', in adverts for perfume or lingerie, where it is almost always a man who is shown buying gifts for his female partner or in medical, scientific or advisory texts (which may focus on male-female penetrative (missionary position) intercourse as the only (or preferred) way of conceiving a child or achieving orgasm). Discourses of compulsory heterosexuality could also be shown by the *absence* of explicit references to heterosexuality in speech and writing, effectively normalizing or unproblematizing the concept. For example, we would expect the terms *man*, *gay man* and *heterosexual man* to occur in general language usage in the order of frequency that I have just listed them in. *Man* is generally taken to mean *heterosexual man*, which is why the latter term would appear so rarely. *Gay man* – being the marked, 'deviant' case would therefore appear more frequently than *heterosexual man*, but not as often as *man*.[1]

Therefore, one way that discourses are constructed is via language. Language (both as an abstract system: phonetics, grammar, lexicon, etc. and as a context-based system of communication) is not the same as discourse, but we can carry out analyses of language in texts in order to uncover traces of discourses.

So bearing this linguistic dimension of discourse analysis in mind, to what extent have corpora been utilized in studies of discourse analysis?

The shift to post-structuralism

Discourse analysts have used corpora in order to analyse data such as political texts (Flowerdew 1997; Fairclough 2000; Piper 2000; Partington 2003), teaching materials (Stubbs and Gerbig 1993; Wickens 1998), scientific writing (Atkinson 1999) and newspaper articles (van Dijk 1991; Morrison and Love 1996; Caldas-Coulthard and Moon 1999; Charteris-Black 2004). Such studies have shown how corpus analysis can uncover ideologies and evidence for disadvantage (see Hunston 2002: 109–23 for a summary).

In addition, corpus-based techniques have been employed in studies which have attempted to analyse differences in language usage based on identity (most notably gender). For example, Shalom's study of men's and women's personal adverts (1997), McEnery *et al*'s (2000) work on swearing and demographic categories in the British National Corpus and Schmid and Fauth's (2003) exploration of gender differences in the ICE corpus. Rey (2001) performed a corpus-based study of dialogue spoken in the television series *Star Trek* looking for differences between male and female language use, while Biber and Burges (2001) looked at changing gender differences in dramatic dialogue using the ARCHER corpus of dramatic texts from the seventeenth to the twentieth century. Holmes (2001) looked at the frequencies of sexist and non-language in a corpus of New Zealand English while Sigley and Holmes (2002) carried out an analysis of frequencies and collocations of the terms *girl(s)* and *boy(s)* in five corpora of British English, concluding that adult females are linguistically constructed as immature with emphasis on their appearance, dependence, domesticity and submissiveness. Finally, Stubbs' (1996) analysis of the ways that gender is constructed within two of Robert Baden-Powell's speeches to boys and girls highlights the fact that ideological issues can be present even around a fairly innocuous word like *happy*. Stubbs showed that Baden-Powell (the founder of the Boy Scouts Association) instructed girls to make other people happy whereas boys were simply instructed to live happy lives.

So while there are a small number of researchers who are already applying corpus methodologies in discourse analysis, this is still a cross-disciplinary field which is somewhat under-subscribed, and appears to be subject to some resistance. Some researchers may acknowledge that theoretically it is a good idea, but continue with mainly qualitative analyses of single texts (or not employ texts at all). Others are more vociferously opposed to corpus-based analysis of discourses. In the process of going to international conferences in various areas of linguistics over the past few years, I have heard interest, disinterest and hostility towards using corpora to analyse discourse in about equal amounts. Part of the problem is perhaps to do with either misconceptions about what corpus analysis actually involves or a dislike of, or unfamiliarity with, computers. Another, more valid issue, which I address below, involves some quite strong (and seemingly incompatible) differences about what counts as 'good' research in both corpus linguistics and discourse analysis. Therefore, it can be difficult to merge both sets of research ideologies.

And while I find corpus-based discourse analysis to be a worthwhile technique, I do not wish to be blindly evangelical about it.

All methods of research have associated problems which need to be addressed and are also limited in terms of what they can and can not achieve. One criticism of corpus-based approaches is that they are too broad – they do not facilitate close readings of texts. However, this is akin to complaining that a telescope only lets us look at faraway phenomena, rather than allowing us to look at things close-up, like a microscope (Partington 1998: 144). Kenny (2001) argues that in fact, the corpus-based approach is more like a kaleidoscope, allowing us to see textual patterns come into focus and recede again as others take their place. Acknowledging what a corpus-based approach can do and what it cannot do is necessary, but should not mean that we discard the methodology altogether – we should just be more clear about when it is appropriate to use it or employ some other method.

Other researchers have problematized corpora as constituting *linguistics applied* rather than *applied linguistics* (e.g. Widdowson 2000). Widdowson claims that corpus linguistics only offers 'a partial account of real language' (2000: 7) because it does not address the lack of correspondence between corpus findings and native speaker intuitions. Widdowson also questions the validity of analysts' interpretations of corpus data and raises questions about the methodological processes that they choose to use, suggesting that the ones which computers find easier to carry out will be chosen in preference to more complex forms of analysis. Additionally, Borsley and Ingham (2002) criticize corpus-based approaches because it is difficult to make conclusions about language if an example does not appear in a corpus. They also argue that language is endowed with meaning by native speakers and therefore cannot be derived from a corpus. See Stubbs (2001a, 2002) for rejoinders to these articles. A related criticism is by Baldry (2000: 36) who argues that corpus linguistics treats language as a self-contained object, 'abstracting text from its context'. And Cameron (1997), in an article about dictionary creation using corpus-based methodologies warns that corpus linguists have had a tendency to over-rely on newspapers and synchronic data, at the expense of charting the historical origins surrounding words and their changing meanings and usages over time. Such criticisms are worth bearing in mind, although should not prevent researchers from using corpora, rather, they should encourage corpus-based work which takes into account potential problems, perhaps supplementing their approach with other methodologies. For example, there is no reason why corpus-based research on lexical items should not use diachronic corpora in order to track changes in word meaning and usage over time and several large-scale corpus building projects have been carried out with the aim of creating historic corpora from different time periods.[2]

Corpus linguistics also tends to be conceptualized (particularly by non-corpus researchers) as a *quantitative* method of analysis: something which is therefore at odds with the direction that social inquiry has taken since the 1980s. Before the 1980s, corpus linguistics had struggled to make an impact upon linguistic research because computers were not sufficiently powerful enough or widely available to put the theoretical principles into practice. Ironically, by the time that computers had become widely available to scholars, there had already occurred a shift in the social sciences in the accepted ways that knowledge was produced via research methodologies. For much of the nineteenth and twentieth centuries, knowledge had been gathered by taking approaches which have been variously called scientific, positivist, essentialist, empirical or structuralist. Such approaches viewed the universe as containing facts or truths that could be discovered by objective researchers working under experimental conditions. They emphasized measurement and categorization – for example, the classification of different species of plants or animals into related groups or the measurement of human characteristics such as height, weight or IQ in order to discover averages or norms. Researchers would form hypotheses and test them under strict experimental conditions. While this approach is often still associated with the natural, physical and biological sciences, it was also used in the social sciences – particularly in sociology, psychology and linguistics where phenomena such as personality, IQ, attitudes and accents were examined.[3] However, social psychologists in the 1960s and early 1970s argued that the discipline was implicitly voicing the values of dominant groups (see Harré and Secord 1972; Brown 1973; and Armistead 1974). Additionally, Gergen (1973) argued that all knowledge is historically and culturally specific and that it isn't possible to look for definitive accounts of people and society, because social life is continually changing. Quantitative research was also criticized as being a form of social regulation in itself (e.g. Hacking 1990) or a way of controlling and predicting (Buchanan 1992), while other researchers (e.g. Cicourel 1964) have argued that quantitative researchers tend to fix meanings in ways that suit their preconceptions.

By the 1980s, an alternative means of producing knowledge had become available, roughly based around the concept of *post-modernism* and referred to as post-structuralism or social constructionism. As Denzin (1988: 432) writes:

> Gone are words like theory, hypothesis, concept, indicator, coding scheme, sampling, validity, and reliability. In their place comes a new language: readerly texts, modes of discourse, cultural poetics, deconstruction, interpretation, domination, feminism, genre, grammatology, hermeneutics, inscription, master narrative, narrative structures, otherness, postmodernism, redemptive ethnography, semiotics, subversion, textuality, tropes.

While Denzin optimistically suggested that now researchers had a *choice* (1988: 432) I would agree with Swann, in her assessment of recent language and gender research, who notes that 'On the whole ... there does seem to have been a shift towards more localised studies' and 'far less reliance is placed on quantifiable and/or general patterns' (2002: 59). So corpus linguistics largely became viable as a methodology at a point where this epistemological shift had already occurred, and its grounding in quantification has not made it attractive to social scientists. Both McEnery and Wilson (1996: 98) and Biber *et al* (1998) also note that the amount of corpus-based research in discourse analysis has been relatively small.

Post-structuralists have developed close formulations between the concepts of language, ideology and hegemony, based on the work of writers like Gramsci (1985) and Bahktin (1984). And the move towards deconstructionism in the social sciences over the past 20 years or so has tended towards research into language and identities that could be particularly associated with people who are viewed as holding or sympathetic towards problematic, contested or powerless identities (for example, gay men and lesbians, women, deaf people, people from non-white ethnic groups, etc). Such people are likely to be aware of the oppression of such groups and therefore hold with forms of analysis that are associated with questioning the *status quo* – e.g. queer theory, feminist linguistics and critical discourse analysis rather than reiterating and reinforcing a list of ways in which people speak, think or behave differently from each other. Burr (1995: 162) refers to this as *action research*, forms of research which have change and intervention rather than the discovery of 'facts' as their explicit aim. Corpus research then, with its initial emphasis on comparing differences through counting, and creating rather than deconstructing categories, could therefore be viewed as somewhat retrograde and incompatible with post-structuralist thinking. Indeed, one area that corpus linguistics has excelled in has been in generating *descriptive* grammars of languages (e.g. Biber *et al* 1999) based on naturally occurring language use, but focusing on language as an abstract system.

Finally, another reason why language and identity researchers have shied away from corpora is due to practical, rather than ideological, considerations. Researchers have argued that discourse analysis is very labour intensive (e.g. Gill 1993: 91) and therefore 'discourse analysis, as with many other varieties of qualitative research is usually *more* difficult than positivist number crunching' (Parker and Burman 1993:

156). However, I would argue that a corpus linguistics approach can be perceived as equally time consuming. Large numbers of texts must first be collected, while their analysis often requires learning how to use computer programs to manipulate data. Statistical tests may be carried out in order to determine whether or not a finding is significant, necessitating the requisite mathematical know-how (or access to a good statistics department). Gaining access to corpora is not always easy – and large corpus building projects can be very time consuming and expensive, sometimes requiring the acquisition of research grants in order to be carried out successfully. No wonder then, that it is often simply less effort to collect a smaller sample of data which can be transcribed and analysed by hand, without the need to use computers or mathematical formulae.

As I stated at the beginning of this section, criticisms of a corpus-based approach are useful in that they make us aware of limitations or potential pitfalls to be avoided. However, having come this far, it seems fair to consider an alternative perspective – what can be *gained* from using corpora to analyse discourse?

Advantages of the corpus-based approach to discourse analysis

Reducing researcher bias

While older, empirical views of research were extremely concerned with the removal of researcher bias in favour of empiricism and objectivity, newer, more post-modern forms of research have argued that the unbiased researcher is in itself a 'discourse of science through which a particular version ... of human life is constructed' (Burr 1995: 160). Burr argues that objectivity is impossible as we all encounter the world from some perspective (the 'objective' stance is still a stance). Instead researchers need to acknowledge their own involvement in their research and reflect on the role it plays in the results that are produced. However, not all discourse analysts are inclined to take this view of objectivity. Blommaert (2005: 31–2) points out, in relation to critical discourse analysis that 'The predominance of biased interpretation begs questions about representativeness, selectivity, partiality, prejudice, and voice (can analysts speak for the average consumer of texts?)'.

It *is* difficult if not impossible to be truly objective, and acknowledging our own positions and biases should be a prerequisite for carrying out and reporting research. However, this perspective assumes

a high degree of researcher self-awareness and agency. The term *critical realism* (Bhaskar 1989) is useful, in that it outlines an approach to social research which accepts that we perceive the world from a particular viewpoint, but the world acts back on us to constrain the ways that we can perceive it. So we need to be aware that our research *is* constructed, but we shouldn't deconstruct it out of existence.

Also, we may be biased on a subconscious level which can be difficult to acknowledge. At other times, we may not want to acknowledge our position for various reasons (concerns, for example, that our findings may be played down because they were published by someone who holds a particular identity, or we may desire to protect or conceal some aspect of our own identity such as sexuality, gender or ethnicity for other reasons). And a lot of academic discourse is written in an impersonal, formal style, so introducing some sort of personal statement may still seem jarring, particularly in some disciplines.

And ultimately, even if we declare our personal circumstances and their relationship to our research, we may still end up being biased in ways which have nothing to do with who we are but are more concerned with the way that human beings process information. A famous study by psychologists Kahneman and Tversky (1973) showed that people (105 out of 152 to be exact) tend to think that in a typical sample of text in the English language the number of words that begin with the letter 'k' is likely to be greater than the number of words that have 'k' as the third letter. In reality, there are about twice as many words that have 'k' as their third letter than there are words that begin with 'k'. Yet we tend to index on the first letter because we can recall such words more easily. We also tend to succumb to other cognitive biases. Mynatt *et al* (1977) showed that in a variety of settings, decision makers tended to notice more, assign more weight to, or actively seek out evidence which confirmed their claims, while they tended to ignore evidence which might discount their claims (*confirmation bias*). Related to this is the *hostile media effect* (Vallone *et al* 1985) which shows that ideological partisans tend to consistently view media coverage as being biased against their particular side of the issue (a phenomenon that perhaps we should attend to when carrying out action research). People also tend to focus more on information that they encounter at the beginning of an activity (the *primacy effect*). The presence of such cognitive biases can be particularly problematic when carrying out discourse analysis. For example, we may select a newspaper article which 'confirms' our suspicions, but ignore other articles which present a different perspective. There is nothing essentially 'wrong' about that, but it may mean that we need to be careful in terms of any generalizations we make beyond the article itself.

Additionally, we may only focus on aspects of a text which support our initial hypotheses, while disregarding those which present a more complex or contradictory picture.

By using a corpus, we at least are able to place a number of restrictions on our cognitive biases. It becomes less easy to be selective about a single newspaper article when we are looking at hundreds of articles – hopefully, overall patterns and trends should show through.

Of course, we cannot remove bias completely. Corpus researchers can theoretically be just as selective as anyone in choosing which aspects of their research to report or bury. And their interpretations of the data they find can also reveal bias. For example, in Chapter 5 of this book I look at the terms *bachelor* and *spinster*, in order to argue that there are strong differences in the ways that meanings and connotations surrounding these words are constructed in a large corpus of general language use. An initial finding in Chapter 5 is that *bachelor* (and its plural form) occur more often than *spinster(s)*. The actual figures are 506 vs. 175. This may lead us to conclude that unmarried men are discussed more in general English than unmarried women, which could be part of a larger trend whereby male terms are more frequent than female terms – an overfocus on men at the expense of women in actual language use. However, a closer look at the data reveals that in some cases *bachelor* actually refers to women. We would also need to take into account the fact that in about 61 cases *bachelor* refers to a type of degree rather than an unmarried man (although we could argue that historically, the two meanings are connected). There are also other cases where *bachelor* refers to proper nouns, e.g. the name of a horse. Again, we may argue that in itself, it is of note that things are named after bachelors but not after spinsters. And we may also decide to focus on words that regularly co-occur with *bachelor*, that tend to index positive attitudes, such as *eligible*, but overlook other, less positive words that also co-occur with *bachelor*, such as *lonely*. With corpus analysis, there are usually a lot of results, and sometimes, because of limitations placed on researchers (such as word length restrictions of journal articles), selectivity does come into play. But at least with a corpus, we are starting (hopefully) from a position whereby the data itself has not been selected in order to confirm existing conscious (or subconscious) biases. One tendency that I have found with corpus analysis, is that there are usually exceptions to any rule or pattern. It is important to report these exceptions alongside the overall patterns or trends, but not to over-report them either.

The incremental effect of discourse

As well as helping to restrict bias, corpus linguistics is a useful way to approach discourse analysis because of the *incremental* effect of discourse. One of the most important ways that discourses are circulated and strengthened in society is via language use, and the task of discourse analysts is to uncover how language is employed, often in quite subtle ways, to reveal underlying discourses. By becoming more aware of how language is drawn on to construct discourses or various ways of looking at the world, we should be more resistant to attempts by writers of texts to manipulate us by suggesting to us what is 'common-sense' or 'accepted wisdom'.

So a single word, phrase or grammatical construction on its own may suggest the existence of a discourse. But other than relying on our intuition (and existing biases), it can sometimes be difficult to tell whether such a discourse is typical or not, particularly as we live in 'a society saturated with literacy' (Blommaert 2005: 108). By collecting numerous supporting examples of a discourse construction, we can start to see a cumulative effect. In terms of how this relates to language, Hoey (2005) refers to the concept of *lexical priming* in the following way: 'Every word is primed for use in discourse as a result of the cumulative effects of an individual's encounters with the word.' As Stubbs (2001b: 215) concludes 'Repeated patterns show that evaluative meanings are not merely personal and idiosyncratic, but widely shared in a discourse community. A word, phrase or construction may trigger a cultural stereotype.' Additionally, Blommaert (2005: 99) notes that a lot of human communication is not a matter of choice but is instead constrained by normativities which are determined by patterns of inequality.

And this is where corpora are useful. An association between two words, occurring repetitively in naturally occurring language, is much better evidence for an underlying hegemonic discourse which is made explicit through the word pairing than a single case. For example, consider the sentence taken from the British magazine *Outdoor Action*: 'Diana, herself a keen sailor despite being confined to a wheelchair for the last 45 years, hopes the boat will encourage more disabled people onto the water.' We may argue here that although the general thrust of this sentence represents disabled people in a positive way, there are a couple of aspects of language use here which raise questions – the use of the phrase *confined to a wheelchair*, and the way that the co-ordinator *despite* prompts the reader to infer that disabled people are not normally expected to be keen sailors. There are certainly traces of different types of discourses within this sentence, but are they typical

or unusual? Which discourse, if any, represents the more hegemonic variety?

Consulting a large corpus of general British English, we find that the words *confined* and *wheelchair* have fairly strong patterns of co-occurrence with each other. The phrase *confined to a wheelchair* occurs 45 times in the corpus, although the more neutral term *wheelchair user(s)* occurs 37 times. However, *wheelchair bound* occurs nine times. We also find quite a few cases of *wheelchair* appearing in connection with co-ordinators like *although* and *despite* (e.g. *despite being restricted to a wheelchair he retains his cheerfulness; despite confinement to a wheelchair, Rex Cunningham had evidently prospered; although confined to a wheelchair for most of her life, Violet was active in church life and helped out with a local Brownie pack*). While this isn't an overwhelmingly frequent pattern, there are enough cases to suggest that one discourse of wheelchair users constructs them as being deficient in a range of ways, and it is therefore of note when they manage to be cheerful, prosperous or active in church life! The original sentence about Diana the keen sailor certainly isn't an isolated case, but conforms to an existing set of expectations about people in wheelchairs. Thus, every time we read or hear a phrase like *wheelchair bound* or *despite being in a wheelchair*, our perceptions of wheelchair users are influenced in a certain way. At some stage, we may even reproduce such discourses ourselves, thereby contributing to the incremental effect without realizing it.

Resistant and changing discourses

As well as being able to establish that repeated patterns of language use demonstrate evidence of particular hegemonic discourses or majority 'common-sense' ways of viewing the world, corpus data can also reveal the opposite – the presence of counter-examples which are much less likely to be uncovered via smaller-scale studies. And if a resistant discourse *is* found when looking at a single text, then we may mistake it for a hegemonic discourse.

Discourses are not static. They continually shift position – a fact that can often be demonstrated via analysis of language change. There is little agreement among linguists about whether language reflects thought or shapes thought or whether the relationship constitutes an unending and unbroken cycle of influence. Whatever the direction of influence, charting changes in language is a useful way of showing how discourse positions in society are also in flux. What was a hegemonic discourse ten years ago may be viewed as a resistant or unacceptable discourse today. At the most basic level, this can

be shown by looking at changing frequencies of word use in a diachronic (or historical) corpus, or by comparing more than one corpus containing texts from different time periods. For example, if we compare two equal sized corpora of British English[4] containing written texts from the early 1960s and the early 1990s we see that in the 1990s corpus there are various types of words which occur much more frequently than they did in the 1960s corpus: e.g.: lexis which reflect the rise of capitalist discourses: *initiatives, strategies, capitalist, customer, resources, privatisation, market*; and lexis which reflect 'green' discourses: *environmental, global, environment, worldwide, conservation*. In addition, we find that certain terms have become less frequent: *girl* and titles like *Mr* and *Mrs* were more popular in 1960s British English than they were in the 1990s, suggesting that perhaps sexist discourses or formal ways of addressing people have become less common.

However, we could also compare the actual contexts that words are used in over different time periods as it may be the case that a word is no more or less frequent than it used to be, but its meanings have changed over time. For example, in the early 1960s corpus the word *blind* almost always appears in a literal sense, referring to people or animals who cannot see. The term *blind* is not significantly more frequent in the 1990s corpus, although in about half its occurrences we now find it being used in a range of more metaphorical (and negative) ways: *turn a blind eye, blind ambition, sheer blind anger, blind panic, blind patriotism, the blind lead the blind, blind to change*. We could say that *blind* has expanded semantically, to refer to cases where someone is ignorant, thoughtless or lacks the ability to think ahead. As Hunston (1999) argues, this non-literal meaning of *blind* could constitute a discourse prosody which influences attitudes to literal blindness (although it could also be argued that the separate meanings exist independently of each other). What the corpus data has shown, however, is that the negative metaphorical meaning of *blind* appears to have increased in written British English over time – it is not a conceptualization which has always been as popular.[5]

Triangulation

As described earlier in this chapter, the shift to post-structuralist methods of thought and research has served to de-emphasize the focus on more quantitative, empirical methods. However, another aspect of post-structuralism may actually warrant the inclusion of corpus-informed research. One of the main arguments of social constructionism is to question and 'deconstruct' binary arguments that have served the

basis of western thinking for thousands of years, such as 'nature or nurture' (Derrida 1978, 1981).

Such oppositions are typical of ideologies in that they create an inherent need to judge one side of the dichotomy as primary and the other as secondary, rather than thinking that neither can exist without the other. Instead, Derrida recommends that we reject the logic of *either/or* of binary oppositions, in favour of a logic of *both/and*. The same could be said for the split between *quantitative/qualitative* or *structuralism/post-structuralism*. Indeed, post-structuralism favours a more eclectic approach to research, whereby different methodologies can be combined together, acting as reinforcers of each other. It is not the case that corpus linguists should view corpora as the only possible source of data; 'Gone is the concept of the corpus as the sole *explicandum* of language use. Present instead is the concept of a balanced corpus being used to *aid* the investigation of a language' (McEnery and Wilson 1996: 169).

Tognini-Bonelli (2001) makes a useful distinction between *corpus-based* and *corpus-driven* investigations. The former uses a corpus as a source of examples, to check researcher intuition or to examine the frequency and/or plausibility of the language contained within a smaller data set. A corpus-driven analysis proceeds in a more inductive way – the corpus itself is the data and the patterns in it are noted as a way of expressing regularities (and exceptions) in language. In this book (apart from in Chapter 7), the case studies I describe are corpus-driven analyses – each one uses a particular corpus as the main or only source of data. However, there is no reason why corpora cannot take more of a corpus-based role in discourse analysis either.

As McNeill (1990: 22) points out, *triangulation* (a term coined by Newby 1977: 123), or using multiple methods of analysis (or forms of data) is now accepted by 'most researchers'. Layder (1993: 128) argues that there are several advantages of triangulation: it facilitates validity checks of hypotheses, it anchors findings in more robust interpretations and explanations, and it allows researchers to respond flexibly to unforeseen problems and aspects of their research. Even when discourse analysts do not want to have to go to the trouble of building a corpus from scratch, they could still gainfully use corpora as a reference, to back up or expand on their findings derived from smaller-scale analyses of single texts (something which I will look at in Chapter 7). For example, Sunderland (2004: 37–8) looked at a newspaper article which publicized a 'fairytale' venue for marriage ceremonies. She argued that the article focused on the bride as the bearer of the (stereotypically) male gaze (due to phrases such as 'its flying staircase down which the bride can make a breathtaking

entrance'). An analysis of the words which *bride* tends to collocate (co-occur) with most often in a large corpus of naturally occurring language revealed terms to do with appearance like *blushing*, *dress*, *wore*, *beautiful* and *looked*. On the other hand, *bridegroom* and *groom* tended to collocate with mainly functional words (pronouns, conjunctions, prepositions, etc.), suggesting that the constructions of brides in the article were 'loaded' in a way which did not apply to bridegrooms. So while the main focus of Sunderland's analysis was a single news article, a general corpus proved to be useful in confirming suspicions that what she was seeing *was*, in fact, a hegemonic discourse. In such cases it only takes a couple of minutes to consult a reference corpus, showing such a corpus-based method to be an extremely productive means of triangulation.

Some concerns

While in the last section I have hoped to show how corpus linguistics can act as a useful method (or supplementary method) of carrying out discourse analysis, there are still a few concerns which are necessary to discuss, before moving on.

First, corpus data is usually only language data (written or transcribed spoken), and discourses are not confined to verbal communication. By holding a door open for a woman, a man could be said to be performing a communicative act which could be discursively interpreted in numerous ways – a discourse of 'the gallant man', of 'male power imposing itself on women' or a non-gendered discourse of 'general politeness in society' for example. In a similar way, discourses can be embedded within images – for example, pictures of heterosexual couples often occur in advertising, helping to normalize the discourse of compulsory heterosexuality, while photo-spreads of women in magazines aimed at (heterosexual or bisexual) men reveal dominant discourses about what constitutes an attractive woman by male standards. Caldas-Coulthard and van Leeuwen (2002) investigate the relationship between the visual representations of children's toys (in terms of design, colour and movement) such as The Rock and Barbie and texts written about them, suggesting that in many cases discourses can be produced via interaction between verbal and visual texts.

The fact that discourses are communicated through means other than words indicates that a corpus-based study is likely to be restricted – any discourses that are uncovered in a corpus are likely to be limited to the verbal domain. Some work has been carried out on creating

and encoding corpora of visual materials, e.g. Smith *et al*'s corpus of children's posters (1998), although at the moment there does not appear to be a standardized way of encoding images in corpora.

In addition to that, issues surrounding the social conditions of production and interpretation of texts are important in helping the researcher understand discourses surrounding them (Fairclough 1989: 25). Questions involving production such as who authored a text, under what circumstances, for what motives and for whom, in addition to questions surrounding the interpretation of a text: who bought, read, accessed, used the text, what were their responses, etc. can not be simply answered by traditional corpus-based techniques, and therefore require knowledge and analysis of how a text exists within the context of society. One problem with a corpus is that it contains decontextualized examples of language. We may not know the ideologies of the text producers in a corpus. In a sense, this can be a methodological advantage, as Hunston (2002: 123) explains '. . . the researcher is encouraged to spell out the steps that lie between what is observed and the interpretation placed on those observations.'

So we need to bear in mind that because corpus data does not interpret itself, it is up to the researcher to make sense of the patterns of language which are found within a corpus, postulating reasons for their existence or looking for further evidence to support hypotheses. Our findings *are* interpretations, which is why we can only talk about restricting bias, not removing it completely. A potential problem with researcher interpretation is that it is open to contestation. Researchers may choose to interpret a corpus-based analysis of language in different ways, depending on their own positions. For example, returning to a study previously mentioned, Rayson *et al* (1997) found that people from socially disadvantaged groups tend to use more non-standard language (*ain't, yeah*) and taboo terms (*fucking, bloody*) than those from more advantaged groups. While the results themselves aren't open to negotiation, the reasons behind them are, and we could form numerous hypotheses depending on our own biases and identities, e.g. poor standards of education or upbringing (lack of knowledge), little exposure to contexts where formal language is required or used (no need to use 'correct' language), rougher life circumstances (language reflecting real life), the terms helping to show identity and group membership (communities of practice), etc. Such hypotheses would require further (and different) forms of research in order to be explored in more detail. This suggests that corpus analysis shares much in common with forms of analysis thought to be qualitative, although at least with corpus analysis the researcher has to provide explanations for results and language patterns that have been discovered in a relatively neutral manner.

Also, a corpus-based analysis will naturally tend to place focus on patterns, with frequency playing no small part in what is reported and what is not. However, frequent patterns of language do not always necessarily imply underlying hegemonic discourses. Or rather, the 'power' of individual texts or speakers in a corpus may not be evenly distributed. A corpus which contains a single (unrepresentative) speech by the leader of a country or religious group, newspaper editor or CEO may carry more weight discursively than hundreds of similar texts which were produced by 'ordinary people'. Similarly, we should not assume that every text in a corpus will originally have had the same size and type of audience. General corpora are often composed of data from numerous sources (newspaper, novels, letters, etc.) and it is likely to have been the case that public forms of media would have reached more people (and therefore possibly had a greater role to play in forming and furthering discourses) than transcripts of private conversations. We may be able to annotate texts in a corpus to take into account aspects of production and reception, such as author occupation/status or estimated readership, but this will not always be possible.

In addition, frequent patterns of language (even when used by powerful text producers) do not always imply mainstream ways of thinking. Sometimes what is *not* said or written is more important than what is there. A hegemonic discourse can be at its most powerful when it does not even have to be invoked, because it is just taken for granted. For example, in university prospectus discourse we would expect to find a term like *mature student* occurring more often than a term like *young student*. However, we should not assume that there are more mature students than young students, as the term *student* implicitly carries connotations of youth and does not need to be expanded upon, hence there is little need for a marked opposite equivalent of *mature student* (*immature student*?). Similarly, a hegemonic discourse can be at its most powerful when it does not even have to be invoked, because it is just taken for granted. A sign of true power is in *not* having to refer to something, because everybody is aware of it. Prior awareness or intuition about what is possible in language should help to alert us to such absences, and often comparisons with a larger normative corpus will reveal what they are.

We also need to be aware that people (as suggested earlier in this chapter) tend to process information rather differently to computers. Therefore, a computer-based analysis will uncover hidden patterns of language. Our theory of language and discourse states that such patterns of language are made all the more powerful because we are not aware of them; therefore we are unconsciously influenced. However, it can be

difficult to verify the unconscious. For example, in Chapter 4 I show how refugees are characterized as out-of-control water, with phrases like *flood of refugees*, *overflowing camps*, *refugees streaming home*, etc. being used to describe them. I (and other researchers) have interpreted this water metaphor as being somewhat negative and dehumanizing. However, would we all interpret *flood of refugees* in the same way? Hoey (2005: 14) points out that we all possess personal corpora with their own lexical primings which are 'by definition irretrievable, unstudiable and unique'. If we were very concerned about the ways that refugees are represented, then we may have already consciously noticed and remarked on this water-metaphor pattern. But what if English was not our first language? Would we be less or more likely to notice and understand the metaphor? And if we were someone who didn't approve of refugees, we may even interpret the word *flood* as being too 'soft', preferring a less subtle negative description. Also, did the person who wrote *flood of refugees* actually intend this term to be understood in a negative sense, or were they simply unthinkingly repeating what has now become a 'naturalised' (El Refaie 2001: 366) way of writing about refugees (as Baker and McEnery 2005 point out, even texts produced by The Office of the United Nations High Commissioner for Refugees, a body aimed at helping refugees, contain phrases using the water metaphor). As Partington (2003: 6) argues, 'authors themselves are seldom fully aware of the meanings their texts convey'. Perhaps conscious intention is more crucial to the *formation* of discourses and reliance on subconscious repetition and acceptance is required for their *maintenance*. See also Hoey (2005: 178–88) for further discussion.

And words do not have static meanings, they change over time. They also have different meanings and triggers for different people. Some people, for example, tend to get annoyed by a recent development of the word *gay* to refer to things that people disapprove of – e.g. 'this exam timetable is so gay' (Baker 2005: 1). However, from talking to people who use the word in this way, many of them do not intend it to be homophobic (some of them are gay themselves) and some (much younger users) are not aware that the word *gay* refers to same-sex attraction or even understand what same-sex attraction is. Corpus analysis needs to take into account the fact that word meanings change and that they can have different connotations for different people.

Therefore, a corpus-based analysis of language is only one possible analysis out of many, and open to contestation. It is an analysis which focuses on norms and frequent patterns within language. However, there can be analyses of language that go against the norms of corpus

data and in particular, research which emphasizes the interpretative repertoires (Gilbert and Mulkay 1984) that people hold in relationship to their language use can be useful at teasing out the complex associations they hold in connection to individual words and phrases.

Corpus linguistics does not provide a single way of analysing data either. As the following chapters in this book show, there are numerous ways of making sense of linguistic patterns: collocations, keywords, frequency lists, clusters, dispersion plots, etc. And within each of these corpus-based techniques the user needs to set boundaries. For example, at what point do we decide that a word in a corpus occurs enough times for it to be 'significant' and worth investigating? Or if we want to look for co-occurrences of sets of words, e.g. how often do *flood* and *refugees* occur near each other, how far apart are we going to allow these words to be? Do we discount cases where the words appear six words apart? Or four words? Unfortunately, there aren't simple answers to questions like this, and instead the results themselves (or external criteria such as word count restrictions on the length of journal articles) can dictate the cut-off points. For example, we may decide to only investigate the ten most frequently occurring lexical words in a given corpus in relation to how discourses are formed. However, while these words tell us something about the genre of the corpus, they may be less revealing of discourses. So we could expand our cut-off point, to investigate the top 20 words. This is more helpful, but then we find that we have too much to say, or we are repeating ourselves by making the same argument, so we make a compromise, only discussing words which illustrate different points.

Again, these concerns should not preclude using corpus data to analyse discourse. But they may mean that other forms of analysis should be used in conjunction with corpus data, or that the researcher needs to take care when forming explanations about her or his results.

Structure of the book

This book has two main goals: to introduce researchers to the different sorts of analytical techniques that can be used with corpus-based discourse analysis, and to show how they can be put into practice on different types of data. Because I feel that people understand better when they are given real life examples, rather than discussing ideas at an abstract level, I have included a range of different case studies in the following chapters in the book (see Table 1.1 for a summary). Chapter 2 looks at issues to do with data collection and corpus building, in order

to address questions such as how large a corpus should be and the best ways to collect and annotate data. Chapter 3 uses a small corpus of holiday leaflets written for young adults in order to examine how some of the more basic corpus-based procedures can be carried out on data and their relevance to discourse analysis. It includes looking at how frequency lists can be used in order to provide researchers with a focus for their analyses and how measures such as the type token ratio help to give an account of the complexity of a text. It also shows how the creation of dispersion plots of lexical items can reveal the development of discourses over the course of a particular text.

Chapter 4 investigates the construction of discourses of refugees in newspaper data and is concerned with methods of presenting and interpreting concordance data. It covers different ways of sorting and examining concordances as well as introducing the concept of semantic and discourse patterns. Chapter 5 uses a large corpus of general British English in order to consider differences in discourses surrounding never-married men and women. Collocations of the words *bachelor* and *spinster* are examined. This chapter explores different ways of calculating collocation and the pros and cons associated with each. It shows how reference corpora can be used to uncover hidden meanings within words or phrases, and how collocational networks can reveal strong associations between central concepts in a text.

Chapter 6 examines different discourse positions within a series of debates on fox-hunting which took place in the British House of Commons. In order to achieve this, we look at the concept of keywords, lexical items which occur statistically more frequently in one text or set of texts when compared with another (often a larger 'benchmark' corpus). However, this chapter expands the notion of keywords to consider key phrases (e.g. multiword units) and key semantic or grammatical categories – which necessitates prior annotation of a text or corpus.

Chapter 7 considers how the corpus approach can be employed in order to examine linguistic phenomena that occur beyond the lexical level, by looking at patterns of nominalization, attribution, modality and metaphor. Using a single news article which contains reference to *allegations of rape*, I use a reference corpus to examine typical language patterns surrounding this term and its related forms in order to show whether the language of the news article is typical or not. Finally, Chapter 8 concludes the book and re-addresses some of the concerns that have been first raised in this chapter.

However, before moving on to look at the different techniques that can be used in order to carry out corpus-based discourse analysis, we

first need a corpus. Chapter 2 therefore explores issues connected to obtaining data, building and annotating a corpus.

Table 1.1 *Texts, topics and methods of analysis used in this book*

Chapter	Text type	Topic	Main Techniques covered
3	Holiday leaflets	Young people/ use of alcohol	Frequency lists, clusters, dispersion plots
4	Newspaper articles	Refugees	Concordances
5	General Corpus	Never-married people	Collocations
6	Political debate	Fox-hunting	Keywords
7	General Corpus	Allegations of rape	Analysis of nominalization, modality, attribution and metaphor

Notes

1. In fact, in the 100 million word British National Corpus, *gay man* appears 17 times, *homosexual man* occurs 6 times and *heterosexual man* appears once. *Straight man* appears 20 times, of which only two occurrences refer to sexuality (the others mainly refer to the 'straight man' of a comedy duo). *Man* (without these sexuality markers) occurs 58,834 times.
2. For example, the Helsinki Corpus of English Texts: Diachronic Part consists of 400 samples of texts covering the period from 750 to 1700 (Kytö and Rissanen 1992).
3. For example, in psychology, researchers had created the notion of different 'personality' traits or scales such as Eysenk's introversion/extroversion scale (1953) which could be quantified via asking subjects a list of questions such as 'Do you enjoy going on roller-coasters?', and then calculating a score based on their answers. However, an extreme social constructionist viewpoint would argue that the concept of personality is unreal because people behave differently in a range of contexts (e.g. depending on whether they are at work or with their parents or different groups of friends). Ironically, psychologists labelled people's ability to adjust their behaviour according to social context as being yet another quantifiable personality trait – *self monitoring* (Snyder and Gangestad 1986). Personality inventories therefore assume that there must be an essential identity, an 'inner me' or true personality, which social constructionists would dispute. In a similar way, Potter and Wetherell (1987: 43–55) questioned the notion of quantitative questionnaire-based 'attitude' research (e.g. Marsh 1976) by carrying out a qualitative analysis of interviews that attempted to elicit attitudes about immigrants. They found that the analyst's categories did not match the participant's terms, elicited attitudes were often contradictory and that defining the status of the object under discussion was problematic.

4. The Lancaster-Oslo/Bergen (LOB) and the Freiberg Lancaster-Oslo/Bergen (FLOB) corpora respectively.
5. The reasons why these changes in uses of *blind* over time have appeared is another matter. Perhaps the more negative idiomatic metaphoric uses of *blind* have always existed in spoken conversation, but were censored in written texts because editors required authors to use language more formally. What is interesting though, is that there has been a shift in written discourse which has resulted in *blind* being conceptualized in a very different way over a 30 year period.