5

Checks and Balances: How Corpus Linguistics can Contribute to CDA

Gerlinde Mautner

Introduction

This chapter focuses on the role that corpus linguistics can play in CDA projects. It will introduce readers to previous work in this area, explain basic concepts and techniques, present two worked examples and encourage critical engagement with the methodology.

Those with previous experience of corpus linguistics will be aware that it is a methodology that uses computer support – in particular, software called 'concordance programs' – to analyse authentic, and usually very large, volumes of textual data. Its potential usefulness for CDA, rather than for lexicography and grammar, may be less familiar, though. Reflections on the potential of combining corpus linguistics and CDA go back quite a long way now (e.g. Hardt-Mautner, 1995), and in the 1997 edited volume on discourse studies (Van Dijk, 1997), de Beaugrande argued that '[l]arge corpuses offer valuable support for the project of discourse analysis to return to authentic data' (de Beaugrande, 1997: 42). Still, none of the other contributors to that edition actually used the method. Awareness of its potential does seem to be growing, however, and there has been a spate of more recent CDA work using corpus linguistics (e.g. Baker and McEnery, 2005; Baker et al., 2007; Baker et al., 2008; Cotterill, 2001; Fairclough, 2000a; Mautner, 2007; Nelson, 2005; Orpin, 2005). Even so, it seems fair to say that the techniques of corpus linguistics are not yet generally regarded as being at the core of CDA's methodological canon. That the present

(second) edition of the volume includes this chapter could thus be said to reflect a change in trend.

What, then, can one expect corpus linguistics to contribute to CDA? In a nutshell, the potential of this methodology rests on three factors:

- Corpus linguistics allows critical discourse analysts to work with much larger data volumes than they can when using purely manual techniques.
- In enabling critical discourse analysts to significantly broaden their empirical base, corpus linguistics can help reduce researcher bias, thus coping with a problem to which CDA is hardly more prone than other social sciences but for which it has come in for harsh and persistent criticism (e.g. Widdowson, 1995, 2004).[1]
- Corpus linguistics software offers both quantitative and qualitative perspectives on textual data, computing frequencies and measures of statistical significance, as well as presenting data extracts in such a way that the researcher can assess individual occurrences of search words, qualitatively examine their collocational environments, describe salient semantic patterns and identify discourse functions.

This chapter cannot offer detailed step-by-step guidance on project design and execution. For that, there are other, and arguably more suitable, sources that readers may want to turn to, notably Baker (2006) and McEnery et al. (2006). However, a few basics will be covered in the following section, using original sample analyses as well as cross-referencing existing work in this area. Throughout, the emphasis will be less on technical detail than on enabling readers to make their own informed judgements on whether the method is right for them. There are two worked examples: the first shows how a large reference corpus can be mined for socially relevant information, establishing a collocational profile of a key expression from the lexis of work, namely *unemployed*. The second takes a single newspaper article as its starting point, and uses large-corpus data as an aid in interpreting what appears to be a particularly 'loaded' expression from the article – the adjective *hard-working*.

Both case studies are based on the assumption that language and the social are inextricably, and dialectally, linked. In other words, the way in which labels, in this case *unemployed* and *hard-working*, are used reflects social attitudes, perspectives and categorizations. And the labels, in turn, shape the way in which social structures and relationships are perceived. By referring to a person or group as *unemployed*, one cannot help implying that being employed is the desired default, just as *hard-working* comes with a host of positive connotations directly related to an essentially capitalist work ethos.

Different approaches to discourse, and concomitant definitions of the term, exist in abundance, as do various notions of what it means to carry out 'critical analysis' (see Wodak, 2004: 198–199 and Wodak, 2006b for comprehensive overviews). The perspective adopted in this chapter is functional and constructivist (as well as unabashedly simple). *Discourse* is taken to refer to authentic texts used

in multi-layered environments to perform social functions. *Analysing* discourse is understood as the systematic attempt to identify patterns in text, link them to patterns in the context, and vice versa. Doing so *critically* means unveiling and challenging taken-for-granted assumptions about language and the social, as well as recognizing discourse as a potentially powerful agent in social change.

It will not have escaped readers' attention that the title of this chapter contains two hedging devices, one modal (*can*) and another lexical (*contribute*). These correspond to two caveats which are worth spelling out right at the beginning. One is that the usefulness or otherwise of this method, as of any other, depends crucially on recognizing what kinds of research questions it is suitable for tackling. With corpus linguistics, the key limiting factors are the capabilities of the software, as well as the features – mainly in terms of composition and annotation – of the electronically held corpora that are used. At the current state of play, and considering the limitations of those tools that are sufficiently widely available, there is a very strong bias in favour of the individual lexical item and clusters thereof. Put simply, 'the word' is the peg that everything else is hung on. It follows that if the linguistic phenomenon you are interested in is in fact tied to, or at least crystallizes around, discrete lexical items, then you are likely to find this method a boon both as a practical and efficient time-saver, and as a powerful heuristic tool helping to clear pathways to discovery. If, on the other hand, the phenomenon to be focused on is one that is played out on a larger textual stage, and with varying and unpredictable lexical realizations, then corpus linguistic methods will be of little or no help. However, at some point or other, as soon as questions of micro-level linguistic realization are addressed, even projects located very much at the macro end of the CDA spectrum will have occasion to benefit from a corpus linguistic approach.

The second caveat, related to the 'contributing' role of corpus linguistics, is that we need to recall one of the principal tenets of what might be termed 'mainstream' CDA (broadly, the traditions shaped by Fairclough, Wodak and Van Dijk[2]), namely that the analyst must, precisely, look *beyond* the text proper in order to unearth socially meaningful interpretations that can then be enlisted to do socially transformative work. We need our much-famed 'context', history, and as firm a grasp as possible of the politics, in the widest sense, that have a bearing on the production and reception of the text. This social hinterland and the textual evidence before us are intricately linked, but rarely in a fully transparent, one-to-one type of relationship; hence the idea of making corpus linguistics 'contribute' to CDA rather than it 'doing CDA' of its own accord. All the same, at an Oscar night of methods, my vote would be on corpus linguistics as Best Supporting Actor, and the present chapter sets out to make the case for that award.

In addition, there is the added benefit that if you decide to include corpus linguistic methods in your CDA project design, you need not in fact discard, 'unlearn' or in any other way throw overboard whatever more traditional methods you have grown accustomed to using. As an ancillary method, corpus

linguistics is flexible and unobtrusive, and if handled appropriately, will enrich but not prejudice the rest of the research design or the interpretation of the results.

Key concepts and a worked example

As with any method, researchers will want to know, first and foremost, what it can do, what kind of data and research questions it is suitable for, and what obstacles may be encountered when applying it. These are the concerns of the present section.

Concordancing software

Programs known as concordancers do not, by themselves, 'produce' analyses, but perform operations on text that make it easier for humans to analyse it. Some of the information that concordancers provide is quantitative, such as absolute and relative word frequencies. Programs also compute measures that indicate the relative statistical significance of the co-occurrence of items. Examples here are t-scores, which capture certainty of collocation, and MI ('mutual information') scores, which tell us about the strength of the bond between two items, that is, whether there is a higher-than-random probability of the two items occurring together (Church and Hanks, 1990; Clear, 1993: 281; Hunston, 2002: 73; McEnery and Wilson, 2001: 86).[3]

To see how this works in practice, let us have a look at the first worked example of this chapter: building a collocational profile of the adjective *unemployed*. Although this serves as a stand-alone case study here, with a focus on method rather than content, it could also form the nucleus of a substantive contribution to the existing body of discourse-analytic research on unemployment (see Muntigl et al., 2000 and the 2002 special issue of *Text*[4]). As we shall see, the corpus linguistic approach allows the researcher to work with enormous amounts of data and yet get a close-up on linguistic detail: a 'best-of-both-worlds' scenario hardly achievable through the use of purely qualitative CDA, pragmatics, ethnography or systemic functional analysis.

Given the key role that the mass media play in constructing social reality, a corpus of newspaper articles would appear to be a suitable starting point. Wordbanks Online,[5] a multi-genre corpus of more than 500 million words of mainly British and American text, includes a nearly 60-million-word corpus of articles from the British daily newspaper *The Times*. This is the subcorpus we will turn to first. The search reveals that *The Times* corpus contains 567 instances of *unemployed*. Table 5.1 gives the ten collocates with the highest t-scores and MI scores respectively.[6]

The t-score part of the table is headed, as is invariably the case, by high-frequency grammatical items. Here, these are *an* and *are*, with *who*, *and* and *for*

TABLE 5.1 *t-scores and MI scores for collocates of* unemployed *in* The Times *corpus of Wordbanks Online*

Collocate	t-score	Collocate	MI score
1. *an*	6.648362	1. *steelmen*	13.465279
2. *are*	6.227450	2. *househusband*	11.780613
3. *people*	5.779392	3. *unemployable*	11.228017
4. *who*	5.725799	4. *housewives*	8.733004
5. *and*	4.842066	5. *4m*	8.066547
6. *term*	4.212151	6. *youths*	7.898282
7. *long*	3.749890	7. *disadvantaged*	7.547531
8. *million*	3.623313	8. *homeless*	7.213343
9. *for*	3.605234	9. *pensioners*	6.965925
10. *workers*	3.516933	10. *claimants*	6.889355

not far behind. Such 'function' words, devoid of separate meaning as they are, tend not to be as interesting to discourse analysts as to grammarians, and it is generally safe in a CDA setting to ignore them and indeed the whole t-score rank scale. Somewhat unusually though, the 'top ten' by t-score here includes five content words (*people, term, long, million* and *workers*), with the presence of both *long* and *term* probably being due to the phrase *long-term unemployed*. Comparing this with the collocation list of two other, randomly picked adjectives, *happy* and *sad*,[7] we can see how unusual it is for content words to appear so high up in the t-score rank scale: the group of top ten in the t-score list for *happy*, for example, contains no content word at all, and the one for *sad* includes only one (the intensifier *very*). For lexical items to be in the same t-score league as grammatical ones points to a high degree of what one might call patterned bonding; that is, of a lexical connection so formulaic that the degree of certainty with which it occurs equals that of patterns involving grammatical items. Translated into CDA terms, the 'phrase-ness' of a noun group referring to people could point to the solidified discursive construction of a social group (a necessary first step towards stereotyping).

Turning now to the right half of Table 5.1, which lists the top ten collocates according to 'Mutual Information' scores, we can see which social attributes being unemployed is associated with: *unemployable, disadvantaged* and *homeless*. Frequent nominal collocates include several labels for marginalized, dependent and economically inactive social groups: *househusband, housewives, youths, pensioners, claimants*.[8] In a full-blown study, rather than one done for demo purposes only, each of these high-frequency collocates would be interesting entry points to the corpus. For example, it would be worth looking at what kind of activities 'unemployed youths' are seen to engage in (by running a search of *unemployed youths* followed by a verb form), how the usage of *househusband* and *housewife* compares, how *unemployed* is linked syntactically to the negative adjectives it frequently collocates with, how quantification (cf. the collocate *4m*) contributes to establishing the unemployed as a problem group, and so on.

These sorts of questions lead us to another feature of concordance programs, one that a discourse analyst with a predominantly qualitative mindset might get more mileage out of than frequencies and statistics: their eponymous capacity to produce concordances. These are extracts from the corpus, displayed in such a way that the search word or phrase (also referred to as the 'node') appears in the middle of a line. The text that the extract comes from can be accessed at all times and with a simple operation such as a double mouse click or selecting an option from a menu bar. Accessible co-texts vary from just over 500 characters (e.g. with Wordbanks Online) to full texts (e.g. with Wordsmith Tools). Lines can be sorted alphabetically: for example, according to the word immediately preceding or following the search word. When sorted like this, the collocational environment of the search word can be assessed rapidly, with frequent patterns standing out clearly. As shown in Table 5.2, for example, the concordance of *unemployed* followed by *and* and another adjective shows up a preponderance of items (six out of a total of nine) with a negative semantic load: *desperate, disadvantaged, divorced, homeless* and *unemployable* (which occurs twice).

TABLE 5.2 *Occurrences of* unemployed, *followed by* and *and another adjective, in* The Times *corpus of Wordbanks Online*

mince. Tony Shalhoub is broke,	unemployed and	**desperate.** His fortune went
<p> Liam Parker, services to	unemployed and	**disadvantaged** people. Alison
near Consett, Co Durham, Sheila,	unemployed and	**divorced,** lives in a council
about 12,000 former teachers are	unemployed and	**free to work.**" <p> He added
new law had effectively made him	unemployed and	**homeless.** He is married with
<p> Ronnie (Ben Miles), 35,	unemployed and	**Jewish,** is back from Israel,
full-time mothers, selfemployed,	unemployed and	**retired** people across the
needed to keep the largely	unemployed and	**unemployable** Saudi young from
the welfare state that bribes the	unemployed and	**unemployable** middle-class

To make sure that this semantic pattern is not just confined to newspaper discourse, it is worth checking the corresponding data from the British spoken corpus. The picture is in fact very similar (Table 5.3).

TABLE 5.3 *Occurrences of* unemployed, *followed by* and *and another adjective, in the British spoken corpus of Wordbanks Online*

to say hello to everybody who's	unemployed and	**bored** at the moment and hasn'
to town looking for work. He was	unemployed and	**homeless** when he turned up at
say he is thirty-two years old,	unemployed and	**single,** and will appear in
apprenticeship's over he becomes	unemployed and	**unemployable** himself. It's

These results are further confirmed if we extend the search to the complete 500+ million word corpus of Wordbanks Online. The adjectives joined to *unemployed* include more occurrences of the items already identified in the two subcorpora (see Tables 5.2 and 5.3) as well as new and similarly negative collocates such as *angry, demoralized, destitute, disabled, dreary, drunk, excluded, poor,*

struggling and *underprivileged*. By examining concordances, therefore, we can see more or less at a glance that the search word, *unemployed*, has a so-called negative 'semantic aura', or 'semantic prosody' (Hunston, 2004: 157; Louw, 1993; Partington, 2004). Alternatively, readers of this volume will be interested to note, this concept has been referred to as '*discourse* prosody', for example by Stubbs (2001: 65), in order to emphasize its role in expressing attitudes and in establishing coherence.

Detractors of corpus-based methods could argue, of course, that one hardly needs a huge database of text and sophisticated software to 'prove' that being unemployed is not a pleasant thing. On the other hand, we should not forget, first, that a fair proportion of any empirical work is devoted, precisely, to finding evidence for the intuitively obvious. Second, some insights may appear 'obvious' *after* having emerged from data but were nothing of the kind before. Supposedly neutral words such as *cause* or *provide*, to use one of Stubbs's examples, can in fact be heavily skewed in terms of their evaluative content when concordance evidence is examined. Patterns are revealed that are not easily accessible even to native speakers' intuition: *cause*, it turns out, collocates predominantly with unpleasant events, such as *damage*, *death*, *disease* or *trouble*, whereas *provide* occurs with desirable things, such as *care*, *help*, *money* or *service* (Stubbs, 2001: 65). Thus, if a speaker or writer uses *provide*, that choice in itself implies that what is provided is being presented as good rather than bad. Third, the concordancer does more than highlight the evaluative polarity, be it good or bad, of an item's collocational environment. Collocates may also turn out to belong to a class of words that share a semantic feature, that is, the search word may have a particular 'semantic preference' (Stubbs, 2001: 88). For example, Baker (2006: 79, 87) concludes from corpus evidence that *refugees* has a semantic preference for quantification, collocating frequently with numbers and phrases such as *more and more* (see also Baker and McEnery, 2005). In a study using a similar approach, Mautner (2007) shows that *elderly* often co-occurs with items from the domains of care, disability and vulnerability. By the same token, if we return to our concordance output related to *unemployed* and its co-ordinated adjectives, we can see that, broadly speaking, these denote either social states (e.g. *available for work, excluded, immigrant, nomadic, unemployable, unpaid*), negative emotions (e.g. *angry, bored, depressed*), or indeed a condition at the interface of both (*unloved*). The collocational profile thus points to the twin nature of unemployment as a social phenomenon with a manifestly psychological impact on individuals.

Taken together, then, semantic preference and discourse prosody show us what kinds of social issues a particular lexical item is bound up in, and what attitudes are commonly associated with it. Importantly, collocational patterns are not merely instantiated in text, but also cling to the lexical items themselves. 'Words which are co-selected', Tognini-Bonelli (2001: 111) reminds us, 'do not maintain their independence. If a word is regularly used in contexts of good news or bad news or judgement, for example, it carries this kind of meaning around with it'.

Finally, some software packages, such as Wordsmith Tools, allow the analyst to compare word lists compiled from different corpora, determining which words are statistically more frequent 'keywords' in a corpus (Baker, 2006: 125; Baker et al., 2008: 278; Mulderrig, 2006: 123). Fairclough's (2000a) study, for example, focuses on the keywords of New Labour – words, that is, that are more frequent in New Labour material than in earlier Labour texts, and more frequent, too, than in general corpora (Fairclough, 2000a: 17).

To sum up, concordancing software offers the above features which are useful for CDA applications (see Table 5.4).

TABLE 5.4 *Tools and types of linguistic evidence provided by concordance software*

Quantitative evidence	Frequency lists
	Comparisons of wordlists, giving information on relative frequency ('keyness')
	Measures of statistical significance: • t-score • 'Mutual Information' (MI) score
Qualitative evidence	Concordance lines sorted alphabetically, enabling the researcher to identify: • semantic preference • semantic prosody

Corpus design issues

These days, when linguists talk about *a corpus*, they generally refer to 'a collection of (1) *machine-readable* (2) *authentic* texts [...] which is (3) *sampled* to be (4) *representative* of a particular language or language variety' (McEnery et al., 2006: 5, original italics). Let us look at each of these four characteristics in turn, with an eye to whatever specific implications they may have for applications in CDA.

Machine readability is the obvious prerequisite for analysing language with the concordancing software described in the previous section. This sounds straightforward enough, but when it is coupled with the second feature – authenticity – issues of data quality arise that critical discourse analysts will want to address. Standard concordancers need 'plain text' files, stripped of formatting, layout and accompanying visuals. While traditional lexico-syntactic research does not see this as a loss, critical discourse analysts will (or should). After all, it is one of the foundational assumptions of discourse analysis, whether of the critical persuasion or not, that meaning-making works simultaneously on several levels, including the non-verbal. Elements of textual design, including typography, colour and text–image relationships, are

not merely embellishments, but play an integral role in making text function as socially situated discourse (van Leeuwen, in this volume). The semiotic reduction that concordancing inevitably entails (Koller and Mautner, 2004) need not jeopardize the validity of one's analyses, but there ought to be adequate safeguards to ensure that whatever is lost along the way can be salvaged at a later stage. In mundanely practical terms, this means collecting and storing hard-copy or scanned originals for future reference, to be drawn upon should multimodality become an issue. Likewise, audio or video recordings of spoken data ought to be preserved, so that contextual clues lost through transcription and conversion into machine-readable format can be retrieved if and when necessary.

As far as criteria (3) and (4) are concerned – sampling and representativeness – the requirements for ensuring methodological rigour are basically no different here than they are for other approaches (see Mautner, 2008). The first step is to identify the 'universe of possible texts' (Titscher et al., 2000: 33), while the second involves sampling. This can be random (i.e. done by first numbering the texts in the 'universe' and then selecting those with the numbers that a random number generator has picked out). Alternatively, it may be guided by criteria that are applied systematically and, in a top-down selection process, narrow down the corpus to a manageable size (e.g. 'take one article about Topic A from newspapers B and C published each week between dates X and Y'). There is a third sampling method, common in qualitative research but unlikely to be suitable for corpus linguistic work, which uses a cyclical process, building a small and homogeneous corpus, then analysing it and adding to it on the basis of the first results (Bauer and Aarts, 2000: 31). The process is repeated until 'saturation' has been reached – a situation, that is, where adding new data does not yield any new representations (Bauer and Aarts, 2000: 34). The problem with this procedure from a corpus linguist's point of view is the flip side of what makes it appealing to the purely qualitative researcher: that you stop collecting data as soon as what you find is simply more of the same. In corpus linguistics, the frequency of an item or structure is taken to be a key indicator of its significance. If you stop adding text to your corpus as soon as repetition becomes apparent, you are effectively closing off any frequency-based line of inquiry. This may well be an acceptable decision to take in a particular project; after all, the qualitative analysis of concordance lines is as important and valuable as the quantitative inquiry that concordancing software allows. But it would have to be a decision taken with full awareness of the loss involved. Certainly, care must be taken not to indulge in hasty judgements about what can be excluded from the corpus on the grounds that it is 'similar' to what is already there. Such rashness can easily defeat the whole purpose of the corpus-building exercise, the point of which is, in a sense, to outwit the analyst who may be tempted to know the data *before* rather than *after* the analysis.

Summing up, corpus design involves the following issues (Figure 5.1):

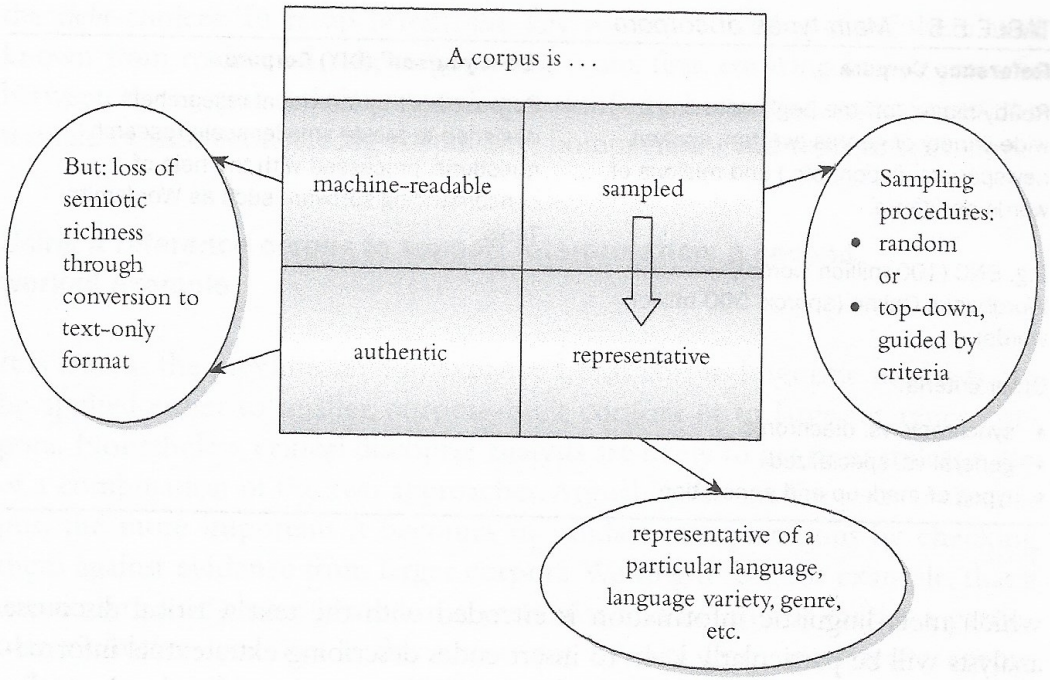

FIGURE 5.1 *Corpus characteristics and issues of corpus design*

Types of corpora and data capture

Corpora come in many shapes and sizes. There are huge, multi-million-word corpora such as The British National Corpus[9] (BNC) and Wordbanks Online, from which the *unemployed* example in the previous section was taken. These are ready-made, commercially available, and each comes with bespoke software unique to it (usually a source of frustration if you want to use both corpora simultaneously). Both were the result of large-scale projects spanning many years and involving teams of linguists and computer experts. In CDA, such corpora are ideal for painting on a very large canvas, investigating how broader social issues are reflected in the genres and discourses represented in the corpus (such as fiction, newspapers and spoken dialogue). This approach is used in studies such as those of Krishnamurthy (1996) on racism and Mautner (2007) on ageism.

At the other end of the spectrum, there are much smaller, 'do-it-yourself' (DIY) corpora (McEnery et al., 2006: 71), purpose-built by individual researchers or small teams to investigate specific research questions. Issues of size apart, corpora may also be classified according to whether they are synchronic, reflecting a language variety at any given point in time, or diachronic, reflecting historical development. They can be general, including a wide variety of genres and media, or specialized, focusing on a particular genre (e.g. corporate mission statements), a particular medium and topic (e.g. web-based texts on disability rights), a particular genre and topic (e.g. parliamentary speeches on global warming), or a particular topic in various genres and media (e.g. the 'Evolution vs. Creationism' debate in sermons, newspaper articles and web logs). Corpora may also differ in terms of

TABLE 5.5 *Main types of corpora*

Reference Corpora	**'Do-it-yourself' (DIY) Corpora**
Ready-made, 'off the peg', including a wide variety of genres (written, spoken, newspapers, fiction, etc.) and millions of words per genre. E.g. BNC (100 million words), Wordbanks Online (approx. 500 million words) Other criteria: • synchronic vs. diachronic • general vs. specialized • types of mark-up and annotation	Purpose-built by individual researchers, designed to tackle smaller-scale research questions, processed with the help of concordancing software such as Wordsmith Tools.

which meta-linguistic information is encoded with the text. Critical discourse analysts will be particularly keen to insert codes describing extratextual information (a procedure known as 'corpus mark-up'), such as text type, speakers' or writers' sociolinguistic characteristics, or indeed any feature that is relevant for a particular research question. Mark-up plays a key role in allowing the analyst to relate the examples shown up by corpus searches back to their original contextual environments (McEnery et al., 2006: 22–23). In addition, corpora may have undergone what is called 'annotation', a process of inserting, for example, parts-of-speech tags, prosodic or semantic information (Baker, 2006: 38–42). Table 5.5 summarizes the main types of corpora.

In building corpora of any size, the World Wide Web has emerged as a key resource. With the exception of spontaneous spoken language (which, admittedly, is a very significant exception), the web offers a huge variety of text and sheer unlimited amounts of it. Cut-and-paste procedures have greatly facilitated the process referred to as 'data capture' (Baker, 2006: 31–38; McEnery et al., 2006: 73). As a result, corpora running to hundreds of thousands of words can be assembled within the space of a few weeks. (On the copyright issues involved, see McEnery et al., 2006: 77–79.) More importantly still, this can be done by individual scholars without the help of battalions of research staff, access to whom is normally dependent on massive research grants and/or a position high up in the institutional pecking order. Seen from this angle, enlisting corpus methods also has a democratizing effect on critical research.

Thus, in terms of sheer corpus availability, most critical discourse analysts' interests will be catered for by the web (which makes it all the more surprising that until very recently, only comparatively few CDA projects were actually based on online material – see Mautner, 2005). Yet in spite of this new abundance, many of the old questions remain, which brings us back to the design issues raised in the previous section. Modern technology may have reduced the manual drudgery of corpus-building, but not the need for brainpower to make

the *right* choices. To recap briefly, the key issues involved – all of them well known from traditional corpus-gathering – are: first, ensuring an adequate fit between the corpus and the research question(s) to be tackled, and, second, the triangle of issues concerning corpus size, homogeneity and representativeness.

Using a reference corpus to support interpretation: a second worked example

It is true, as the previous section explained, that corpus linguistic methods can be applied either to smaller, purpose-built corpora or to large reference corpora. Nonetheless, critical discourse analysts are likely to get most mileage out of a combination of the two approaches. Arguably, the smaller your own corpus, the more important it becomes to validate interpretations by checking them against evidence from larger corpora. We might feel, for example, that a certain word or phrase is used in a particular text because it carries a particular evaluative load, but is this borne out by data from a reference corpus reflecting general language use? What kind of collocational 'baggage' do words occurring in the text carry in wider universes of discourse? It is through comparative evidence from large corpora that suspicions can be confirmed or rejected, thus safeguarding against 'overinterpretation and underinterpretation' (O'Halloran and Coffin, 2004). It has to be conceded, of course, that making sense of this comparative evidence still involves acts of interpretation on the part of the analyst. Neither the quantitative nor the qualitative evidence that large corpora lay before us speaks for itself and to claim that it does so would be seriously misguided or, at the very least, naive. On the other hand, surely, any improvement in CDA's empirical credentials ought to be welcome, tempered though it may be by the sobering realization that completely mechanized discourse analysis is impossible. Or, were it possible, it would cease to be critical.

The idea of comparative evidence is illustrated by the second worked example. It relates to a column in the British popular daily the *Sun* (25 July 2007), published at a time when major floods were devastating large areas in central England. The headline reads 'It's time to turn off the spongers' money tap' (with *spongers*, or *scroungers*, referring to people who receive money but do nothing in return). The article is built around the contrast between the lives of 'hard-working people' whose homes have been destroyed by the flood in Gloucestershire, and a family with 12 children from Berkshire (*the scrounging Gillespies*) who have been given a new council house paid for by *the hard-working taxpayer*. The father of the family is quoted as saying, 'if it was economical for me to work then I would do'. A benefits system producing such attitudes, the article demands, ought to be changed; also, foreign aid should be put on hold 'until *our* national crisis is sorted out' (original italics). There are two large pictures, one of a flooded house in Gloucester, and a second of the Gillespies' £500,000 council home, which has a

small inset showing the couple surrounded by their 12 children. Addressing politicians from both government and opposition, the columnist pleads: 'The hard-working people of Britain should come first'.

Like many opinion pieces in the popular press, this column develops a stark black-and-white dichotomy, pitting 'us' against 'them'. 'Us' refers to the *grafting taxpayer* (with *grafting* being a colloquial British expression meaning 'hard-working'), whose *hard-earned home* has been flooded, whereas 'they' are *the blatantly idle Gillespies of this world* and, less prominently, recipients of development aid in Africa. In establishing the 'we' group, the adjective *hard-working* clearly plays a central role. In an article of around 1300 words, it occurs five times. In four of these instances, *hard-working* is part of a noun group which includes at least one other in-group marker (*people of Britain, British taxpayers, our own people, the communities*). The occurrences, quoted below, are spread fairly evenly throughout the text, contributing to its cohesion:

1. ... there's only one basic rule to remember: the **hard-working** people of Britain should come first.
2. Yet frankly, when the **hard-working** British taxpayers need them most, our politicians look as washed up as the million-plus doormats floating around the streets.
3. ... not to mention the indirect £8 billion of our taxes the PM [Prime Minister] has recently pledged to Africa. But when our own **hard-working** people are in trouble, there should be an instant amnesty on all other benevolent activity until *our* national crisis is sorted out [original italics].
4. There's more rain on the way, so I suggest Brown [= the Prime Minister], Cameron [= the Leader of the Opposition] et al. get their galoshes on and *show* they care about the blighted, **hard-working** communities (...) [original italics].
5. ... the skewed thinking that the **hard-working** taxpayer is a cash cow only to be milked and never fed.

On this evidence, it would be fair to claim that in this text, *hard-working* does ideological work, establishing a 'we' group and attributing a positive quality to it. But then, what is it about *hard-working*, exactly, that makes it such a powerfully positive label? And is this evaluative load true of general language use, or specific to certain discourses? In other words, is there a specifically 'tabloid' attitude towards 'hard-working' people?

Let us see what the Wordbanks reference corpus has to say. First of all, the collocation list for *hard-working* from the total 500+ million-word corpus, ordered by 'mutual information' score, reveals a long list of other positive adjectives. Picking out those that also have double-digit joint frequencies (that is, not only exhibit a strong collocational bond with *hard-working*, evident through an MI score of five or above, but also occur at least ten times), we arrive at the following list (see Table 5.6).

TABLE 5.6 *Collocates of* hard-working *in the total 500+ million-word Wordbanks corpus, with joint frequencies of at least 10 and MI scores of at least 5*

Word	Joint frequency	MI score
industrious	11	9.531038
conscientious	19	9.158034
abiding	19	8.612447
honest	75	7.322699
dedicated	46	7.307235
disciplined	11	7.087851
loyal	28	7.047991
sincere	12	7.001116
competent	12	6.847253
ambitious	28	6.793461
decent	38	6.724341
intelligent	27	6.611245
enthusiastic	14	6.242391
caring	17	6.127628
talented	15	6.044615
skilled	10	6.039344

Some of these are fairly closely related to the idea of working hard or in a particularly focused manner (*industrious, conscientious, dedicated, disciplined, ambitious, skilled*), but others refer to very general attributes that are quite independent of the domain of employment: *honest, loyal, sincere, decent* and *caring*. There is a semantic preference for character traits, and the semantic aura is unequivocally positive. To find that there is a statistically significant collocational bond between these adjectives and *hard-working* means that when someone is described as 'hard-working', there is a higher than random possibility that one of these other, non-work-related qualities will appear in close proximity. For each of these collocates, of course, we need to check what the syntactic relationship with *hard-working* is, because 'being close' could also mean 'close and linked through *but*', in which case the other adjective would be expressing a contrast, not a confirmation, of the virtues implied in hard work (cf. the hypothetical, not attested, phrase *hard-working but caring*). However, it can be established easily by examining the relevant concordances – such as the one for *decent*, which is given in Table 5.7 – that the 'virtuous' adjectives are linked to *hard-working* through *and* or a comma and do in fact refer to the same individuals or groups of people. In addition to showing how *hard-working* and *decent* (highlighted in bold capitals) are related syntactically and semantically when they appear together, the following concordance also reveals a number of other positive attributes in close proximity (highlighted in bold), some with distinctly moralizing overtones (e.g. *honest family man, genuine and Christian, self-sacrificing*).

Such instances of collocation – repeated, statistically significant and attested across a multi-million-word corpus – 'provide objective, empirical evidence for evaluative meanings', and these meanings 'are not merely personal and

TABLE 5.7 *Co-occurrence of* hard-working *and* decent *in the 500+ million-word Wordbanks Online corpus*

to denigrate Kilbane, who is a	hard-working	and **DECENT** professional. Yet on
a more **responsible, DECENT** and	hard-working	British citizen. **He is a credit to**
is that they are **DECENT** and	hard working	people," he said. Mr Xynias said
n the barrel, most fathers are	hard working	**DECENT family types** trying to do
thousands of **DECENT, brave,**	hard-working	coppers. It's hardly surprising
that Hart was a **DECENT,**	hard-working	and **honest family man,** but added
lost on Saturday night. **DECENT,**	hard-working	people, people who are prepared to
things are bad when a **DECENT,**	hard-working	father resorts to taking surgeons
s just as you see him. **DECENT,**	hard-working,	**genuine** and **Christian.** He's the
The bishops are **DECENT,**	hard-working	men in thankless roles, but the
Paul Duckworth was a **DECENT,**	hard-working	and **loving father.** <p> "The
The former teacher is **DECENT,**	hard-working	and **dutiful.** But as my colleague
of him is that he isa **DECENT,**	hard-working	bloke who was caught up in
TV.<p> Meanwhile most **DECENT,**	hard working	citizens will be lucky to see any
great to her family -a **DECENT,**	hard-working	girl." <hl> Open house contest
town I grew up in was a **DECENT,**	hard-working,	hard-drinking, cloth-cap- and-
Erfurt, Germany, as "a **DECENT,**	hard-working	man". <p> But Judge Gareth
he said: "They were all **DECENT,**	hard-working	men - **great lads and great mates.** I
make life a misery for **DECENT,**	hard-working	people. <p> The phone number to
badly on thousands of **DECENT,**	hard-working	taxi men and women who want these
employer as '**good, DECENT,**	hard-working	men" who were mown down in a hail
with admiration as **DECENT,**	hard-working	people, who despite having very
but my parents were **DECENT,**	hard-working	people. We used to go to church
of the game - the **DECENT,**	hard-working	people who work in and around
When I see a **DECENT,**	hard-working	man like you, with a responsible
admired his wife, a **DECENT,**	hard-working,	**self-sacrificing** woman; he couldn'
coal miners, each **DECENT,**	hard-working	union men with large families,
coal miners, each **DECENT,**	hard-working	union men with large families,
Slick Willy" into a **DECENT,**	hard-working	child of the middle class. He told
with New York: the **DECENT,**	hard-working	people who live here. And here's
he said, "a generally	hard-working	and **DECENT** people prepared to put
found them **DECENT, kindly,**	hard-working,	and **knowledgeable** within their
they will effectively stop many	hard working	**DECENT** Sikhs from earning a
of humanity; there's a lot of	hard-working,	**DECENT** people, a lot of children
level-headed, reliable,	hard-working,	**DECENT, orderly"** - are
and more time listening to the	hard-working,	**DECENT** majority that elected New
character who came from a "very	hard-working	and **DECENT** family". <p> The judge
<p> The puzzling aspect is why	hard-working,	**DECENT** people can see what is

idiosyncratic, but widely shared in a discourse community' (Stubbs, 2001: 215). Effectively, the frequent collocates of a word *become* part of its meaning. Thus, by drawing on corpus-based collocational information, a discourse analyst can replace his or her individual, intuitive judgement on evaluative meaning with shared assumptions and judgements.

In further attempting to put the use of *hard-working* in this *Sun* article into perspective, another angle worth looking at is to see how it is used in a newspaper catering for a different readership. The relevant subcorpus that Wordbanks Online offers is the nearly 60 million words from the British daily *The Times*. Whereas

more than 60 per cent of the *Sun*'s readers are in the C2, D and E social grades, 89 per cent of *The Times'* readers belong to the A/B/C1 socio-economic group.[10] The results (Table 5.8) show that *hard-working* occurs more than twice as often in the *Sun* (in relative terms, that is, per million words) than it does in *The Times*.

TABLE 5.8 *Frequency of* hard-working *in the* Sun *and* Times *subcorpora of Wordbanks Online*

	absolute frequency	relative frequency, per one million words
Sun	393	8.69
Times	243	4.06

Furthermore, although the lists of high-frequency collocates appear to be rather similar in the two subcorpora, containing many of the items that showed up when we examined the whole 500-million-word corpus (see Table 5.6), two differences between the *Sun*'s and *The Times'* collocation lists do stand out. One is that *honest* and *decent*, though present in both lists, are relatively more frequent in the *Sun* corpus.[11] The other is that the collocation list for *hard-working* in the *Sun* includes a collocate – the one with the highest MI score, in fact – that is not present in *The Times'* list at all: *abiding*. Switching from the collocation list to concordance mode, we can see that all six occurrences of *abiding* are due to *law-abiding* being one of the positive attributes closely associated with *hard-working* (Table 5.9).

TABLE 5.9 *Co-occurrence of* hard-working *and* law-abiding *in the* Sun *corpus of Wordbanks Online*

against the respectable,	hard-working,	**LAW-ABIDING** majority. No wonder
redit to their **LAW-ABIDING** and	hard-working	community. <p> A V DAVAR, East
>Then hopefully **LAW-ABIDING**,	hard-working	parents like the Gells need neve
t better meals than a lot of	hard-working	**LAW-ABIDING** people, better medic
time all decent, **LAW ABIDING**,	hard-working	people were given some
even exists. <p> How many	hard-working, honest,	**LAW-ABIDING** people can

The social construction involved in this collocational link, rather than any 'objective' semantic association, could hardly be more obvious. After all, it is perfectly possible to be 'lazy' and abide by the law, or work a very busy 70-hour week dedicated to breaking it. Incidentally, taken by itself, *law-abiding* is also considerably more frequent in the *Sun* (145 instances, or 3.2 occurrences per million words) than in *The Times* (111 instances, or 1.86 per million).

Summing up, and relating the evidence back to the article that made us turn to the reference corpus for support, we can draw the following conclusions:

- *Hard-working* is much more than a descriptive label. Its semantic preference and prosody, evident in the concordance lines, indicate that it is part and parcel of a moralizing discourse, linking hard work with positive attributes such as decency, honesty, loyalty, family values and the like.

- These patterns are relatively more prominent in the popular, working-class tabloid the *Sun* than they are in *The Times*, which caters for a predominantly middle-class readership.
- In a critical discourse analysis of a text using *hard-working*, a good case can therefore be made for arguing that the contribution of *hard-working* to the overall meaning of the text is based partly on the ideological baggage that the word carries, and that this, in turn, is derived from attested patterns of usage in larger universes of discourse.

Large-corpus evidence thus provides 'checks and balances' by opening a window on values and attitudes present throughout a discourse community rather than held only by individual researchers.

Summary and critique

Corpus linguistics has a lot to offer to CDA. It helps researchers cope with large amounts of textual data, thus bolstering CDA's empirical foundations, reducing researchers' bias and enhancing the credibility of analyses. On the other hand, critical discourse analysts ought to be self-confident enough to assert that, conversely, corpus linguistics is enriched by being applied to research questions inspired by social concerns, such as power, inequality and change. Ultimately, through their 'theoretical and methodological cross-pollination' (Baker et al., 2008: 297), both CDA and corpus linguistics ought to benefit.

Combining the two approaches typically involves the following steps:

- compiling an electronically held corpus that allows the investigation of research questions arising from social issues
- running the corpus through concordancing software that compiles frequency lists, identifies keywords and reveals statistically significant collocations
- analysing concordances qualitatively in order to establish the dominant semantic preferences and prosodies of lexical items relevant to the social issues under investigation
- putting the results from the purpose-built corpus into perspective by comparing them with evidence gleaned from large reference corpora.

Alternatively, a multi-million-word reference corpus may itself serve as the starting point, allowing researchers to build collocational profiles of socially contested lexical items across a wide range of genres, media and geographical areas.

In spite of the clear benefits involved, there are some areas of potential concern, which I will deal with in turn under five headings (Figure 5.2).

1. The skills gap and lack of standardization

This is a practical rather than a substantive issue, and may well disappear over time. At the time of writing, though, it still looms rather large. To anyone advocating

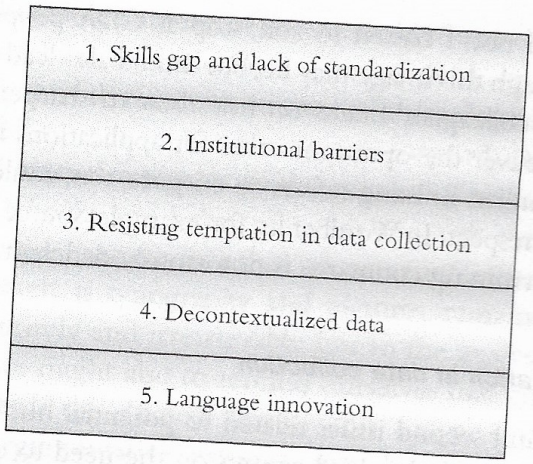

FIGURE 5.2 *Using corpus linguistics in CDA: key areas of concern*

the integration of corpus linguistics into mainstream CDA, it is quite tempting to downplay the effort involved and make reassuring noises along the lines of 'it's not rocket science'. Indeed it isn't, but there is no denying the fact that becoming a confident user takes time and effort. Not much, perhaps, for the mundane task of learning to master the tools; but certainly a significant amount in order to develop the type of mindset that can appreciate the potential of the method, recognize its limitations, hone your analytical skills and refine your discovery procedures, so that ultimately you are able to fashion your research designs accordingly.

The continuing reluctance of many discourse analysts to become involved may well be due in part to the deplorable lack of standardization within corpus linguistics. The British National Corpus and Wordbanks Online, to use just two examples of multi-million-word corpora, do not use the same software. The same is true of the various concordancing packages available for analysing DIY corpora (such as Wordsmith Tools or Monoconc Pro). Search commands differ, screens differ, analytical tools differ: not a happy state of affairs if all you want to do is get on with the job.

2. Institutional barriers

The second point is related to the first, but located on the institutional rather than the individual level. Critical discourse analysts and computer linguists do not necessarily work in the same departments and, if they do, may not communicate well with each other. They often go to different conferences and publish in different journals. As a junior researcher, you are likely to be socialized into either the one methodology or the other, but rarely into both. As most linguists know, but not all care to admit, it is often early exposure to a particular methodology, rather than any inherent merits this may have, that tends to bias one's methodological choices for a long time.

At the risk of launching into after-dinner-speech mode, this is the moment to call for more communication between critical discourse analysts and computer

linguists. This should not, I hasten to add, stop at CDA people begging for IT support, realistic though this image may be, but should also lead to corpus linguists picking their CDA colleagues' brains on how best to sharpen their computing tools so that they deliver the optimum value for applications in socially relevant, applied discourse studies. Existing reference corpora, too, could profit from some overhauling in that respect. In Wordbanks, for example, source referencing – such a key factor in determining context – is notoriously deficient.

3. Resisting temptation in data collection

Whereas the first and second issues related to potential hurdles encountered by those new to the method, the third centres on the need to curb the enthusiasm of the newly converted. We saw earlier that the World Wide Web and electronic processing have made for temptingly laden data tables. And indeed, being able to assemble and analyse large corpora is a key element in defusing the 'cherry-picking' charges frequently levelled at CDA. Generally speaking, corpus size undoubtedly boosts representativeness, and this, in turn, enhances the validity of analysts' claims. On the other hand, as is so often the case, a technological advance comes with strings attached. Somewhat paradoxically, the ease with which corpora can be assembled can prove to be at once overwhelming and tempting for the analyst, novice and seasoned researcher alike. They may well react like a glutton at an all-you-can-eat diner, guzzling data 'food' indiscriminately without due regard for the principles of discerning composition, be it of a menu or of a corpus. In our case, these will revolve, as ever, around questions such as: what kinds of texts are most likely to allow me to answer my research questions? Is the selection of texts which make up my corpus reasonably representative of the 'universe of discourse' that is 'out there'? None of these questions, and the principles underlying them, have ceased to be relevant. If anything, they have become more pressing, precisely because of the *embarras de richesses* surrounding the analyst engaged in corpus-building. Amid the bewildering surplus of easily storable text, it has become easier to lose sight of the need for constant reflexivity, even in the early stages of a project, and particularly with regard to what should go into the corpus. This is not a plea for allowing too much biased selectivity too soon; if it were, it would amount to reverting to the very 'cherry-picking' procedure that a corpus-based approach wishes to counteract (and which is why cyclical corpus-building up to saturation was rejected in our discussion of sampling). The point is, rather, that critical discourse analysts putting corpora together should, quite simply, not get carried away.

4. Decontextualized data

The fourth area of concern, mentioned earlier but worth restating here, relates to the fact that both the input to and output from concordancing software is decontextualized, semiotically reduced language. Although programs allow instant access to wider co-texts or even the full texts that the concordance lines

come from, a considerable amount of non-verbal information is lost when text is transferred to machine-readable form. Corpus mark-up can help compensate up to a point but, with the current state of technological development and commercial availability, it is impossible to run concordancing software while preserving the full textual integrity of the original. This is an area, therefore, where the idea of 'checks and balances' needs to work the other way round, with the analyst having to make sure that whatever information concordancers cannot deal with, such as typography and pictures, remains accessible somewhere and is not entirely and irretrievably lost to the analysis.

In this context, we ought also to remind ourselves that concordancing software is biased towards the discrete lexical unit. Larger-scale discursive phenomena, such as argumentative patterns, may be captured through corpus linguistic techniques, but only if they crystallize systematically around certain words, phrases or lexico-semantic patterns.

Finally, and precisely because corpus linguistics has this fantastic potential for focusing on linguistic detail, there is a need to guard against becoming so engrossed in building collocational profiles of ideologically loaded individual words that the bigger picture is lost. There is a fine line between an eye for detail and myopia. Returning briefly to the *hard-working* example in the previous section: were this part of a full-blown study, it would of course be insufficient merely to look at *hard-working* (central though it is to this newspaper article). In addition, one would not only have to explore the full range of synonyms and related expressions that the article uses, but would also have to delve deeper into the history, politics and social psychology of wage labour and the work ethic.

5. Language innovation

The fifth issue to be borne in mind is that large and static reference corpora, such as the BNC and Wordbanks Online, are useless for investigating developments at the sharp end of language. Where social change is at its fastest – and arguably of keenest interest to CDA – these corpora fall silent. Youth culture, advertising and code-switching varieties emerging among new immigrant populations would be cases in point. For such applications, building ad-hoc DIY corpora is the only solution.

Essentially, the last four areas of concern all relate to the same issue: the need for a realistic assessment of a method's potential. Our metaphorical tools for setting to work on text are subject to very much the same limitations as tools in a literal sense. It makes as much, or as little, sense to criticize corpus linguistic methods for not permitting more contextually embedded analysis, or a static, ten-year-old corpus for being silent on the latest neologisms, as it does to criticize a screwdriver for being no good at hammering in nails.

What is more, it can be tempting to expect computer-based methods to work miracles and to obscure or even compensate for flaws elsewhere in the research design. Put more bluntly, if your sampling technique is faulty and your sample skewed, it will remain so even when computerized, concordanced and

subjected to every statistical procedure under the sun. Similarly, if your choice of statistical techniques is such that the data are, as Baker (2006: 179) graphically puts it, 'subtly "massaged"' in order to produce the desired results; if the analyst reports results selectively, or ignores inconvenient concordance lines, then the fault lies not with their methodology but with their integrity.

Whatever the limitations of corpus linguistics, the complexity of discourse is such that any change in perspective and any insight not otherwise available ought surely to be welcome as additions to the methodological toolbox. On the other hand, we should not forget that, in choosing methods, there is a rather thin dividing line between, on the one hand, eclecticism that is imaginative and productive, and, on the other, aimless patchworking, which is neither. Whether your research design ends up on the right side of this divide depends crucially on (1) a clear statement of the aims of your project, (2) a rigorous assessment of what each method can and cannot do, and (3) robust theoretical foundations capturing core assumptions about language and the social. If deployed wisely, corpus linguistics provides an enriching complement to qualitative CDA, aiding discovery and adding analytical rigour. To return to the metaphor introduced earlier: even an Oscar-winning supporting actor cannot rescue a bad film, but they can make a good film great.

FURTHER READING

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

This book is ideally suited for critical discourse analysts who are first-time users of corpus linguistic methods. It combines theoretical background with hands-on advice and several worked examples.

McEnery, T., Xiao, R. and Yukio Tono, Y. (2006) *Corpus-based Language Studies: An Advanced Resource Book*. London and New York: Routledge.

This book caters for both novice and more experienced researchers, proceeding from the basics to a section with key readings from corpus-based language studies. In a third section, six extended case studies are presented, covering areas as diverse as pedagogical lexicography, L2 acquisition, sociolinguistics, and contrastive and translation studies.

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

Stubbs, M. (2001) *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford and Cambridge, MA: Blackwell.

Both volumes are seminal classics. Although the author positions them, as their respective subtitles reveal, as 'Computer-assisted Studies of Language and Culture' and 'Corpus Studies of Lexical Semantics', rather than CDA, they make essential reading for critical discourse analysts, and especially those who may be 'challenged' in terms of empirical rigour.

Acknowledgement

Material from the Bank of English® reproduced with the kind permission of HarperCollins Publishers Ltd.

Notes

1 This criticism, in turn, has been refuted strongly by, for example, Fairclough (1996) and Wodak (2006a: 606–609).
2 See Fairclough (1992a); Fairclough (1995a); Fairclough and Wodak (1997); Toolan (2002); van Dijk (2007a); Wodak (2006b); Wodak and Chilton (2005).
3 The MI score relates the *observed* frequency of a given co-occurring item within a certain collocational span to the left and right of the search word to the *expected* frequency of the co-occurring item in that span (McEnery et al., 2006: 56). For details of the statistical computation involved, see Matsumoto (2003: 398–399).
4 The issue in question is *Text* 22(3). See, in particular, Graham and Paulsen (2002); Muntigl (2002a); Wodak and van Leeuwen (2002).
5 See www.collins.co.uk/books.aspx?group=154
6 The cut-off point above which measures are considered to indicate statistical significance is 2 for t-scores and 3 for MI scores (Hunston, 2002: 71–72).
7 On the basis of the same number of occurrences, 567, as in the *unemployed* example.
8 In this particular corpus, *steelmen* is not as promising a collocate to follow up as it looks, because all nine occurrences refer to the film *The Full Monty*.
9 See www.natcorp.ox.ac.uk/
10 According to figures from the National Readership Survey, available at www.nrs.co.uk, accessed 9 August 2007.
11 In a million words, there are 0.3 occurrences of *honest* in the *Sun* (15 instances) and 0.1 in *The Times* (seven instances). The figures for *decent* are 0.26 per million for the *Sun* (12 instances) and 0.06 for the *Times* (four instances).