5 *Applications of corpora in applied linguistics*

The application of corpora expected to have most relevance to readers of this book – language teaching – is dealt with in chapters of its own (chapters 6–8). In this chapter, other applications are described. These are:

- The production of dictionaries and grammars, that is, reference books for language learners and translators.
- The use of corpora in critical linguistics, illuminating items of importance to the study of ideologies.
- The use of corpora in translation.
- The contribution of corpora to literary studies and stylistics.
- The use of corpora in forensic linguistics.
- The use of corpora in designing writer support packages.

Dictionaries and grammars

Introduction

Corpora have so revolutionised the writing of dictionaries and grammar books for language learners (or rather, for learners of English) that it is by now virtually unheard-of for a large publishing company to produce a learner's dictionary or grammar reference book that does not claim to be based on a corpus. As a result, this is probably the application of corpora that is most far-reaching and influential, in that even people who have never heard of a corpus are using the product of corpus investigation. Accounts of using corpora to write dictionaries are found in Sinclair (ed.) 1987; Summers 1996; Baugh et al 1996; Clear et al 1996.

This section will concentrate on those areas in which the use of corpora has changed dictionaries and other reference books. These can be summarised as a series of new emphases:

- an emphasis on frequency;
- an emphasis on collocation and phraseology;
- an emphasis on variation;
- an emphasis on lexis in grammar;
- an emphasis on authenticity.

Each of these will now be dealt with in turn.

Emphasis on frequency

One area in which speaker intuition is acknowledged to be of very little use is in the assessment of relative frequency between words, meanings and usages. A very important impact of corpora upon dictionaries, therefore, is the inclusion of information about relative frequencies. Such information may be given explicitly to the dictionary user, or it may be used by the dictionary writer in deciding, for example, which sense of a word to show first.

Because a corpus can show the diversity of use, and the importance, of very frequent words, current learner's dictionaries tend to include more detailed information than the old ones did about these words. A rough indication of this can be given by comparing the number of senses given for certain frequent words in different dictionaries.[1] Here are some comparisons between the *Longman Dictionary of Contemporary English* 2nd edition (1987), which was written without a corpus, and two dictionaries written with the aid of a corpus: the *Longman Dictionary of Contemporary English* 3rd edition (1995) and the *Collins* COBUILD *English Dictionary* (1995).

KNOW

Longman 1987 gives 20 senses of *KNOW*. Longman 1995 gives over 40 and COBUILD 1995 gives over 30.

MATTER

Longman 1987 gives 10 senses of the noun and verb *MATTER*, including phrases such as *as a matter of fact*. Longman 1995 gives over 30 senses. COBUILD 1995 gives over 20.

MAY

Longman 1987 gives 7 senses of the modal *may*. Longman 1995 gives 8 senses. COBUILD 1995 gives 15 senses.

PLACE

Longman 1987 gives 20 senses of the noun *PLACE*. Longman 1995 and COBUILD 1995 each give over 30 senses.

[1] A comparison between number of senses can only be a very rough guide to comprehensiveness because dictionaries do not divide information between senses in consistent ways. One dictionary may include a lot of information in a single sense, where another dictionary may choose to divide the information between two or more senses.

Many of the increases in number of senses is explained by more information being given about the very frequent uses, involving the division of one sense into two or more. For example, Longman 1987 gives as one sense of *matter* the meaning 'something wrong', as in *What's the matter?* Longman 1995 expands this to four senses: the question asking about illness or a state of being broken; the question asking about feelings; the statement *there's something the matter with*; and the negative statement *there's nothing the matter with*.

There are also 'new' senses, that is, meanings or uses that seemed unimportant before a corpus showed how frequent they were. One example is the phrase *I know*, which was included in Longman 1987 only as a phrase uttered when someone gets a sudden idea (as in *What can we get her for her birthday? Oh I know, we'll get her some flowers*). In Longman 1995, two additional uses are given: agreement (as in *'I'm so worn out.' 'Yeah, I know'*) and forestalling disagreement (as in *It sounds silly I know, but try it anyway*). COBUILD 1995 gives three uses of *I know*: agreement; prefacing a disagreement (as in *'There are trains straight from Cambridge.' 'I know, but it's no quicker'*); and showing sympathy and understanding (as in *I know what you're going through*). Another example is *may*. One meaning of *may* in Longman 1987 is the 'degree of certainty' meaning. Longman 1995 divides this into a 'future' meaning (something may or may not happen) and a 'present' meaning (something may or may not be true). COBUILD 1995 adds another sense which indicates a degree of usuality rather than a degree of certainty, that is, something is certain but is only sometimes true (as in *Up to five inches of snow may cover the mountains*). Finally, *place* with a possessive (*my place, your place*) is noted by Longman 1995 and COBUILD 1995 to indicate 'the house or flat where someone lives'. In addition, COBUILD 1995 notes that *the place* is used anaphorically, referring to somewhere that has already been mentioned in the discourse. Neither of these uses of *PLACE* were given in Longman 1987.

Another innovation in dictionaries that has been made possible by the use of a corpus is the inclusion of explicit frequency information. COBUILD 1995, for example, places all the words in the dictionary in one of 6 frequency bands. Longman 1995 notes words that are particularly frequent in spoken and written English and compares the frequency of some words in the two modes. The verbs *BET, MEAN* and *THINK*, for example, are shown to be much more frequent in spoken than in written English, because of the colloquial use of phrases such as *I bet, I mean* and *I think*. *NEED* is shown as more frequent than *REQUIRE* in both written and spoken English, but the difference is greater in spoken English, indicating that *REQUIRE* is more frequently used in written English than in spoken.

Using tagged corpora, it is possible to compare the frequency of the same word, for example, as a noun and as a verb. The dictionary definitions can then be presented in a sensible order, with the most frequent use first (Summers 1996: 262). For example, *GORGE* as a noun (meaning 'a valley with steep sides') is approximately four times as frequent as the verb *GORGE* (meaning 'eat greedily'), so it is sensible to put the noun sense before the verb sense in a dictionary. What would help lexicographers even more would be automatic sense recognition (Clear 1994; Summers 1996), which would enable senses to be identified in terms of frequency also. Given that the different senses of words have different collocations and patterning, automatic sense differentiation is in theory possible, but has not yet been achieved.

Frequency and grammar reference books

There are two ways of dealing with information about frequency in grammar books. One approach is simply to focus on usages that are relatively frequent. Sinclair et al (1990), for example, take this approach. More recently, some writers of grammar books give precise statistical information based on frequency counts in specially designed and annotated corpora (Biber et al 1999; Mindt 2000). Much of this information is to do with variation, and examples are given in the section below on 'Emphasis on variation'. One very interesting type of frequency that is given is the distribution of meanings across a given form. Mindt (2000: 224), for example, identifies four meanings of the present perfect (the indefinite past, past continuing into present, the recent past, and a use indicating that an action is completed, though at an unspecified time).[2] Of these, the first (indefinite past) accounts for almost 80% of all occurrences of the present perfect, with the second (past continuing into present) accounting for all but 5% of the others. The 'recent past' and 'completed action' meanings are comparatively rare. This is in conflict with many course books, which teach uses such as *I have lived here for 12 years* (past-into-present) or *they have recently had their third child* (recent past) as prototypical, when in fact they are less common than the indefinite past use. Most of the other tense

[2] Mindt (2000) in fact divides the indefinite past use into two: resultative and non-resultative. Of these, the resultative use, where the present situation is a result of a past action, is more frequent.

forms presented by Mindt have similarly asymmetrical patterns of use.

This kind of work is probably most useful when frequency can be linked to discourse. For example, Biber et al (1999) point out that the most frequent use of *the* in academic prose is at the beginning of complex noun phrases (such as *the disorientating effect of zero gravity*). Also, the most frequent use of *this* and *these* in the same register is to refer back to ideas previously mentioned. Knowledge of this raises awareness of the use of *the* and *this/these* in academic writing such as the following:

> During the past year, three Danish engineers working at the Technical University of Denmark . . . have been studying a detailed model of *the* side-to-side motion of train wheels. Like all models of train wheels, theirs has a single left/right symmetry – reflectional symmetry about a line perpendicular to *the* axle that connects the wheels. *The* equations of motion that describe *this* system have reflectional symmetry too. But *the* solutions to *these* equations may or may not be symmetric – that is, the wheels need not stay 'centred' on the rails. (*New Scientist*)

For the most part, though, frequency information of this kind is more useful to the syllabus designer or coursebook writer than to the class teacher. For example, according to Biber et al (1999: 388–389), the verb *TELL* is most frequently found in the pattern 'verb + indirect object + complement clause' (e.g. *You can't tell her to get off*) whereas *PROMISE* is most frequently found in the pattern 'verb + complement clause' (e.g. *They promised to write*). *PROMISE* occurs fairly frequently as an intransitive verb (e.g. *I promise*), whereas *TELL* very rarely does (e.g. *time will tell*). If a language course consists of fairly large quantities of authentic language, it is likely that this proportion of frequencies will be mirrored in that language. If only small quantities are used, or if invented language is presented to the learner, frequency information such as this can be used to ensure that the more frequent patterns are presented earlier and more frequently than the infrequent ones.

The limitation of frequency information of this kind is that it can suggest that very infrequent uses can legitimately be ignored. For example, Mindt (2000: 182) reports that 98% of verbs in the past tense refer to past time, which is hardly surprising. References to a hypothetical future, for example, are very rare indeed. It seems safe to assume, therefore, that hypothetical meanings are unimportant from the point of view of the past tense, and learners can safely not be taught this meaning. However, it is important for learners to learn to express the hypothetical, and for this the past tense is significant; in some contexts (such as following *what if*) the past tense is the most frequent. In other words, a student learning the past tense can safely ignore hypothetical meaning, but for a student learning to express hypotheticality the past tense cannot be ignored.

In addition, lexical variation needs to be taken into account when considering frequency (as in the example of *TELL* and *PROMISE* above). For example, Mindt (2000: 185) reports that 'quasi-subjects' such as *it* and *there* are very infrequent with lexical verbs, compared with other subject types. This observation masks a considerable amount of lexical variation. There are a fairly large number of verbs (see Francis et al 1996: 518–542 for lists) for which the sequence '*it* + verb + clause' is very important. The point is that a teacher should not be misled into thinking that the sequence is unimportant just because it occurs frequently only with some lexical items.

Emphasis on collocation and phraseology

The attention paid to phrases such as *I know*, *your place* and *there's something the matter with* in dictionaries written using corpora reflects the tendency of a corpus to highlight collocation and phraseology (Hanks 1987; Sinclair 1987a; 1987b; Summers 1996). For example, in Longman 1987, one sense of *brink* was given as 'as far as one can go without being in a condition or situation'. In Longman 1995, the word by itself is not defined but the phrase *be on the brink of* is defined as 'to be almost in a new and very different situation', with the example *Karl is on the brink of a brilliant acting career.* COBUILD 1995 includes the phraseology in the definition: 'If you are on the brink of something, usually something important, terrible, or exciting, you are just about to do it or experience it.' This definition includes reference to the emotive nature of the situation you might be on the brink of. The examples given reflect the 'central and typical' use of *on the brink of* by indicating a bad situation rather than a good one: *Their economy is teetering on the brink of collapse . . . Failure to communicate had brought the two nations to the brink of war.* The examples illustrate frequent collocates of *the brink of*: the preposition *on* and the verb *TEETER*, and the preposition *to* with the verb *BRING*. (See Sinclair 2000 for a more complete study of *brink*.)

This example illustrates several characteristics of the 'new' dictionaries:

- the tendency, where possible, to define a phrase rather than a word: *be on the brink of* rather than *brink*;

- the use by some dictionaries of the definition sentence to illustrate phraseology;
- the possibility of introducing further collocational information into the definition: 'usually something important, terrible, or exciting';
- the use of examples to introduce more information about collocation.

It also illustrates a dilemma of the dictionary writer: how to be selective from a wealth of information. The word *brink* is used in a variety of phrases, of which *on the brink of* is the most frequent. The most frequent verb before this phrase is *BE*, but *TEETER*, *STAND* and *be poised* are also common, as is *HOVER*. The second most frequent phrase is *to the brink of*, preceded by verbs such as *BRING*, *TAKE*, *DRIVE* and *PUSH*. All these phrases are typically, but by no means invariably, followed by nouns indicating something bad. Examples such as *Roy Evans is guiding them to the brink of a glorious new era* or *was on the brink of promotion* are unusual but not 'wrong'. The problem is that there is too much information here to be dealt with in a brief dictionary entry. Longman 1995 and COBUILD 1995 both choose the most frequent phrase for their definitions. COBUILD 1995 tries to deal with the other phrases in examples, hoping that the learner will extrapolate what is essential (the phrases *on the brink of* and *to the brink of*) and what is useful (the verbs *teetering* and *has brought*).

Phraseology is particularly important in the case of very frequent words, the majority of whose uses may be in fairly fixed phrases (Sinclair 1987b; 1999; Summers 1996). Summers, for example, notes that *day* most frequently occurs in phrases such as *one day*, *the other day* and *some day* (1996: 262–263). Sinclair (1999) argues that dictionaries sometimes do not go far enough in identifying such phraseology, particularly in relation to grammatical, as opposed to lexical, words. He points out, for example, that many instances of the word *a* are accounted for by phrases such as *come to a head*, rather than occurring as an alternative to *the* or another determiner, but that dictionaries (and grammar books) rarely record such usage.

Emphasis on variation

In a keynote lecture at the 2nd North American Symposium on Corpora and Language Teaching (2000), Susan Conrad argued that reference books must cease to be 'monolithic', that is, must cease to treat English as a single entity. Instead, Conrad advocated the approach adopted in the *Longman Grammar of Spoken and Written English* (Biber et al 1999), in which grammatical features are presented in terms of a comparison of frequency between four broadly defined 'registers': fiction, academic prose, 'news' (newspaper reportage), and conversation. Each section of Biber et al (1999) includes corpus evidence of comparative frequency, together with an interpretation which relates the figures to contexts of use. For example, instead of simply describing the formation and use of present and past tense in English, Biber et al (1999) note that present tense occurs more frequently than past tense in conversation and in academic prose whereas past tense occurs more frequently than present tense in fiction. News uses both tenses about equally (1999: 456). These figures are interpreted in terms of the typical meanings made in each register. Conversation frequently uses present tense, for example, because of 'speakers' general focus on the immediate context' (1999: 457). For academic writing, the reason for the preponderance of present tense is different: there is a concern with 'general truth', in which specific time is not relevant (1999: 458). Past tense is typical of narrative, which makes up most of fiction.

A similar approach is taken by Mindt (2000), though he uses only three registers: conversation, fiction and expository prose. As an example of his results, he notes that passives with *BE* are most frequent in expository prose and least frequent in conversation (2000: 269). For passives with *GET*, however, the reverse is true: these forms are most frequent in conversation and least frequent in expository prose (2000: 282). Passives with *BECOME* are most frequent in fiction and least frequent in conversation (2000: 282).

Work of this kind raises the question as to whether the registers selected for comparison are sufficiently homogeneous, or whether they themselves are open to the charge of being monolithic. The news texts used by Biber et al, for example, come from a variety of newspaper types (tabloid and broadsheet, for example) and from different parts of the newspapers (international news, arts reviews, business reports and so on). It is not possible to tell whether certain grammatical features are more prevalent in some of these newspapers or parts of newspapers than others. Similarly, their academic corpus is made up of both academic articles and books, from a range of different disciplines (Biber et al 1999: 31–33). Research reported in chapter 8 of this book, which relates grammatical features to the specific concerns of individual disciplines, suggests that the notion of 'academic prose' as a single register might be an overly blunt instrument. Other questions arise regarding the usefulness of this

comparative quantitative information to the teacher; these will be discussed in chapter 7.

Emphasis on lexis in grammar

Another striking feature of the *Longman Grammar of Spoken and Written English* is the degree to which lexical information forms an integral part of the grammatical description. For example, part of the study of present and past tenses in English includes lists of verbs which are overwhelmingly found in present tense (such as *BET, DOUBT, KNOW, MATTER, MEAN, MIND, RECKON, SUPPOSE* and *THANK*) and those which are more frequent in past tense (e.g. *EXCLAIM, EYE, GLANCE, GRIN, NOD, PAUSE, REMARK, REPLY, SHRUG, SIGH, SMILE* and *WHISPER*) (Biber et al 1999: 459). In this concern for lexis the writers concur with Sinclair, who, in fact, rejects the distinction between lexis and grammar (see chapter 6). The *Collins* COBUILD *English Grammar* (Sinclair et al 1990) was a pioneer in this respect. Describing the imperative, for example, it notes the use of this verb-form in sentences such as *Consider, for example, the contrast between the way schools today treat space and time* which focus the reader's attention on a particular aspect or example of the topic being explained (p204). A list of verbs used in this way is included: *compare, consider, contrast, imagine, look at, picture, suppose* and *take*. (Note that Sinclair et al take a rather different view of the importance of frequency than do Biber et al. Whereas Biber et al tend to give the most frequent verbs to be found with a certain form in a given register, Sinclair et al list the verbs which are most important to a given meaning made in a particular way.)

Sinclair et al (1990) in fact contains many lists, including lists of verbs with particular complementation patterns (p139–193), lists of nouns followed by prepositions (p131) and a list of adjectives used after a noun (p75). What emerges from many of these lists is the fact that words with similar behaviours tend to have similar meanings. For example, nouns followed by *for* mostly indicate a reaction or feeling towards someone or something: *admiration, disdain, dislike, love, regard, respect* or *sympathy*. The feeling is often one of intense wanting: *appetite, craving, desire, hunger, need*, and *thirst*. Another meaning is 'looking or asking for': *bid, demand, quest* and *search*.

The association between pattern and meaning provides the basis for a larger COBUILD project: the Grammar Patterns series (Francis et al 1996; 1997; 1998, see also Hunston et al 1997; Hunston and Francis 1998; 1999). These books are based on the grammar codings

used in the *Collins* COBUILD *English Dictionary* (CCED) (1995), in which a sequence of codes illustrates the sequence of elements in a pattern. Here are some examples of the verb *DECIDE*, together with their pattern coding (taken from the CCED):

- *She decided to do a secretarial course.* The verb is followed by a to-infinitive clause. The coding is V to-inf.
- *He has decided that he doesn't want to embarrass the movement . . .* The verb is followed by a that-clause. The coding is V that.
- *The house needed totally rebuilding, so we decided against buying it.* The verb is followed by a prepositional phrase beginning with *against*. The coding is V *against* n.
- *Its outcome will decide whether Russia's economy can be reformed at all.* The verb is followed by a clause beginning with a wh-word. The coding is V wh.
- *Think about it very carefully before you decide.* The verb is intransitive; it is not followed by anything. The coding is V.

A pattern is identified from a corpus, rather than from a single example. To qualify as a pattern, a phrase or clause or word must frequently occur with the node word and must be dependent on it, as was discussed in chapter 3 with relation to the noun *SUGGESTION*.

When the words that share a pattern are listed, most of them can be seen to share an aspect of meaning. For example, many verbs with the pattern V n *to* n (as in *conceded victory to the ruling party*) have something to do with 'giving', e.g. *accord, administer, allocate, allot, arrogate, assign, award, bequeath, bring, cede, commit, concede, contribute, dedicate, delegate, deliver, devolve, dispense, distribute* and so on. Other groups with the same pattern include:

- verbs to do with communication: *address, admit, announce, bid (farewell), break (news), commend, communicate, confess, confide, describe, dictate, disclose, divulge* and so on;
- verbs to do with ascribing a quality to someone or something: *ascribe, assign, attach, attribute, credit, impute, put down*;
- verbs to do with change: *abbreviate, change, commute, convert, decrease, demote, drop, increase, lower, promote, raise, reduce, shorten, swell, turn, cut down, narrow down, whittle down*;
- verbs to do with devoting yourself to something: *abandon, address, apply, commit, confine, dedicate, devote, enslave, limit, pledge, restrict, rivet, tie, give (oneself) over*;
- verbs to do with adding something to something: *add, affix, annex, append, attach, bind, bolt, chain, clamp, clip, connect, couple* and so on;

- verbs to do with attracting someone: *attract*, *commend*, *draw*, *endear*, *recommend*;
- verbs to do with moving someone from one job or position to another: *accredit*, *appoint*, *apprentice*, *demote*, *nominate*, *ordain*, *promote*, *recall*, *recruit*, *relegate*, *transfer*, *upgrade*;
- verbs to do with betrayal: *betray*, *denounce*, *report*, *shop*, *grass up*, *turn in*;
- verbs to do with changing awareness or attitude: *acclimatise*, *accommodate*, *accustom*, *adapt*, *adjust*, *alert*, *awaken*, *blind*, *desensitise*, *inure*, *reconcile*, *resign*, *sensitise*;
- verbs to do with directing attention: *direct*, *divert*, *draw*, *give (thought)*, *pay*, *switch*, *turn*.

(For complete listings, see Francis et al 1996: 417–433. Comparable corpus-based studies of pattern and lexis are Rudanko 1996; Levin et al 1997.)

Reference books of this kind emphasise the connection between meaning and pattern and provide a resource for vocabulary building in which the word is treated as part of a phrase rather than in isolation. They also provide evidence for a challenge to the traditional distinction between lexis and grammar (see chapter 6), and indeed challenge our view of what a grammatical description is. To illustrate this, consider how grammar is used to account for a particular instance of language. Here is an example from a letter to a newspaper:

> I consider myself to be a so-called 'new man' because I gave up a profession to bring up our son. My wife went back to work a year ago, and since then I have been astonished to discover how many women consider bringing up a baby is a woman's job. I meet a lot of women who find it incomprehensible that I run the household. It's time that more women changed their attitudes – or they can never hope to change those of their menfolk!

Possible grammatical approaches to this text might include examining the tense usage, or the balance between material and mental processes, or the level of modality. A lexical approach to grammar, however, would see it in terms of the patterns belonging to each of the various lexical items, as shown in Table 5.1.

Emphasis on authenticity

When reference books are written with the aid of a corpus, examples can be chosen that illustrate the most typical use of a word or phrase and, if examples are taken from the corpus itself, authenticity is guaranteed, in the sense that each example has been used in genuine

Table 5.1. Grammar patterns in a short text

Text	Pattern
consider myself to be	V n to-inf
be a so-called 'new man'	V n
gave up a profession	V P n
bring up our son	V adv prep
went back to work	v-link ADJ to-inf
have been astonished to discover	V wh
discover how many women consider	V that
consider bringing up a baby is a woman's job	V P n
bringing up a baby	V n
is a woman's job	V n
meet a lot of women	V *it* ADJ that
find it incomprehensible that	V n
run the household	*it* v-link N that
It's time that	V n
changed their attitudes	V to-inf
hope to change	V n
change those of their menfolk	

communication. It is important to recognise that authenticity and typicality are not the same thing. Any corpus contains numerous examples of a given word that are authentic – they are part of actual texts – but are the product of innovation, word-play, or simple odd circumstances, and are therefore not typical. Conversely, some dictionary writers invent sentences that reflect typical usage but which have not been used in authentic situations.

Although all dictionary writers agree that typicality is important, they do not all agree that absolute authenticity is desirable. Baugh et al (1996: 43) argue that:

> Most citations are unsuitable for a learner dictionary because they are too complex grammatically, contain unnecessary difficult words or idioms, or make culture-dependent allusions or references to specific contexts.

The introduction to Longman 1995 (xvi) says:

> All the examples in this dictionary are based on what we find in the spoken and written corpus material in the Longman Corpus Network. Some examples are taken direct from the corpus; some have been changed slightly from the corpus to remove difficult words; and some have been written specially for the entry.

The *Cambridge Learner's Dictionary* (2001: 5) stresses naturalness and typicality rather than authenticity:

> The corpus . . . helps us find natural and typical examples to show how words and phrases are used.

COBUILD 1995 (xxii) makes the strongest claim to authenticity itself:

> The majority of the examples in the dictionary are taken word for word from one of the texts in the Bank of English. Occasionally, we have made very minor changes to them, so that they are more successful as dictionary examples.

Fox (1987) argues the case for authentic examples, pointing out that invented examples often do not reflect nuances of usage. The phrase *take aback*, for example, is typically used in the passive (*I was taken aback by . . .*) rather than in the active, whereas an invented example may be of the type *His reaction took me aback* (though not, of course, if the writer has observed the data accurately). Invented examples, which do not make reference to specific contexts, are often over-explicit. Fox illustrates this with the example *'We'll try to salvage your leg,' said the doctor to the trapped man*, which sounds stilted because in a narrative of which this sentence was a part it would be unlikely that both the doctor and the man would need to be identified at this point. (On the other hand, it might be argued that a less stilted version – *'We'll try to salvage your leg,' he told him* – is less informative and so less helpful.) Fox also indicates the pitfalls of shortening sentences for inclusion in a dictionary. She cites the example *His anguish was terrible*, which comes from an actual sentence reading *His anguish was terrible for her to behold*. Fox comments:

> What seems to be wrong with the short version is that it is too bald and also comes too close to stating the obvious – we naturally assume that anguish is terrible. Perhaps more importantly, in the short version the word 'terrible' describes 'his anguish', whereas in the original it is more plausibly interpreted as referring to her reactions to his grief. (Fox 1987: 148)

Summary of reference books discussion

Reference books for learners of English, then, have been transformed by the use of corpora in their compilation and have become, on the whole, even less like similar books for native speakers. They are greatly influenced by the ease with which information on frequency and typicality is obtained from a corpus, or from contrasting corpora, and tend to emphasise phraseology and the interaction

between lexis and grammar. Whilst there is disagreement on the extent to which examples should be authentic in the sense of 'have been said or written', there is a concern for idiomaticity and realism in examples.

Studying ideology and culture

Ideology in a specialised corpus

A growing concern in Applied Linguistics is the relation between language and ideology, in particular, the role of language in forming and transmitting assumptions about what the world is and should be like, and the role of language in maintaining (or challenging) existing power relations. The dominant school of research into language and ideology is critical linguistics (Fowler 1987) or critical discourse analysis (Fairclough 1995). Critical linguistics looks at language, not as a system on its own but as something that 'intervenes' in the social world, largely by perpetuating the assumptions and values of that world (Fowler 1987: 482–3). Fowler mentions three aspects of what critical linguists do that are important when considering their use of corpora.

1. Critical linguists study texts in the context of the social circumstances in which they have been produced.
2. Critical linguists aim to reveal 'the ideology coded implicitly behind the overt propositions'.
3. Critical linguistics 'challenges common sense by pointing out that something could have been represented some other way, with a very different significance'.

It is clear that the techniques of corpus investigation have much to offer the second and third of these objectives. Patterns of association – how lexical items tend to co-occur – are built up over large amounts of text and are often unavailable to intuition or conscious awareness. They can therefore convey messages implicitly and even be at odds with an overt statement. The different options open to speakers can be illustrated by the various ways that individual words are used. For example, a process indicated by the word *change* can be expressed as a noun, as in *One change that has begun to emerge within the republic over the last ten to twenty years has been a revival of Islam*, or as a verb, as in *people who were trying to change society*. In the noun example, the process of change has no agent, that is, no-one is made responsible for the change. But this is not an inevitable way of expressing what has happened. In the verb example, *people* are

shown as responsible for social change (see Fairclough 2000: 32–34 on metaphors associated with *change* as a noun).

With respect to the first objective – to study texts in their social context – the role of a corpus is less clear. If a corpus is composed of a number of texts, corpus search and processing techniques, such as word-lists, concordance lines and lists of collocations, will tend to obscure the character of each text as a text. Each individual example is taken out of context – that, in a sense, is the point. Furthermore, the corpus treats texts as autonomous entities: the role of the text producer and the society of which they are a part tends to be obscured. Perhaps for this reason, some critical linguists have avoided using corpora, though Krishnamurthy (1996), Stubbs (1997), Caldas-Coulthard and Moon (1999), Piper (2000a) and Mautner (2000) are among those who argue for the value of corpora in such studies.

In this section we look at the role of corpora in critical linguistics in the light of a number of studies. These are:

- Teubert's (2000) study of the language of Euroscepticism in Britain, based on a corpus of texts downloaded from web-sites taking an antagonistic stance regarding the European Union.
- Flowerdew's (1997) discussion of speeches by the last British governor of Hong Kong, Chris Patten. Flowerdew argues that Patten created through these speeches a mythical picture of Britain as a benevolent colonial power.
- Fairclough's (2000) investigation of the language of New Labour, the re-modelled version of the Labour Party that became the party of government in Britain in 1997 after many years in opposition.
- Piper's (2000a, b) analysis of a key concept in the New Labour programme – that of *lifelong learning* – based on a corpus of texts downloaded from British government and EU web-sites.
- Morrison and Love's (1996) study of letters to Zimbabwean magazines ten years after independence.
- Stubbs and Gerbig's (1993) comparison of textbooks in Geography and Ecology, particularly with relation to their use of ergative verbs.
- Wickens' (1998) account of computer-aided teaching materials in Law, which he compares with seminars and textbooks in the same subject.

Most of these writers acknowledge the influence of Stubbs (1996), whose work is discussed further in the next section.

Teubert's (2000) study is one of those that focuses on 'keywords' in the sense used by Williams (1976, cited in Stubbs 1996), that is,

words that have a particular significance in a given discourse (though not identified statistically as described in chapter 4). In Teubert's case some of the words are identified intuitively as conceptually significant in some texts, while others occur as the collocates of those words. They include what Teubert, following Hermanns (1994), calls 'stigma keywords', which indicate an adversary, such as *bureaucrat*, *corruption*, and, in the context of Europe, *federal*, and 'banner keywords', which indicate a positive value, such as *independence*, *peace* and *prosperity*. Teubert notes a considerable amount of repetition between the texts forming his corpus, with phrases such as *bureaucratic dictatorship* and *signing away . . . rights* appearing in several different texts. He relates this repetition to the minority stance of the writers: 'It is this tight net of intertextual references that is indicative of groups basing their identity on an ideological foundation not shared by their environment' (2000: 53). Teubert also points out the density of co-occurrence of the stigma and banner keywords in his texts. Out of 17 examples of the sentences containing the stigma words *unelected* and *faceless*, for example, ten include other stigma words, such as *bureaucrats*, *unaccountable* and *dictators*. A similar pattern is repeated across the other keywords discussed.

The contrast between stigma and banner keywords allows Teubert to draw attention to the inconsistencies in the Eurosceptics' position. For example, *unaccountable bankers* are evidence of the perfidy of Europe, whereas an *independent central bank* is held up as an ideal, yet both *unaccountable* and *independent* indicate institutions which do not answer to a political power (2000: 55). Some words are identified as the site of conflict. *Anglo-Saxon*, for example, is quoted in the corpus as a term of abuse (*nasty Anglo-Saxon capitalist excesses . . . detested Anglo-Saxon model . . . dreadful Anglo-Saxon inequality*), but this use is ascribed to others, and the writers in the corpus themselves use the term positively (*the ideals of Anglo-Saxon democracy . . . Anglo-Saxon economies with their greater flexibility*). The term becomes a rallying point, delimiting quite effectively those for whom it is a stigma term and those for whom it is a banner word (2000: 66).

Teubert uses the identification of recurrent items, phrases and collocations to unpack the assumptions behind the Eurosceptic discourse and to make explicit what is implied but left unsaid: that, according to the Eurosceptics, only Britain out of the whole of Europe is a true democracy with a truly accountable government (2000: 76–77). In doing so he reveals the 'subliminal message' conveyed by the repetition of lexical items and the formation of pattern of association.

Flowerdew (1997; see also Flowerdew 1998) reveals an analogous hidden message in Patten's speeches, arguing that 'lexical reiteration and patterning' is important in suggesting (or 'creating the myth') that Western ideas about the market economy, freedom of the individual, the rule of law, and democratic participation had a benign influence on Hong Kong during its period as a British colony. Evidence for this argument again comes from collocational information. For example, in Patten's speeches the words *economy* and *economic* are usually found in positive environments. Typical collocates are *choice*, *freedom*, *fairness*, *cheerfulness*, *growth*, *good health*, *virtues*, *benefits*, *positive change*, *success*, *talent* and *initiative*. These words not only create a prosody of 'goodness', but also link *economy* to other Western values such as *choice* and *freedom*. Similar points are made about the words *wealth*, *individual* and *rule of law*. Here are Flowerdew's selected examples illustrating Patten's use of the words *individual*, *individuals* and *individuality*:

> the individual against the state
> the individual against the collective
> the rights of the individual
> decency and fairness, individuality and enterprise
> respect for the individual's rights
> the individual's rights to privacy
> individuals and families free to run their own lives
> opportunities for individuals to shape their own lives
> the privacy of individuals
> individuals and their right to seek the protection of the courts
> the freedom of individuals to manage their affairs without fear of arbitrary [illegible]

The words that co-occur with *individual* here are positive ones: *rights*, *respect*, *free*, *opportunities*. *Individuality* co-occurs with *decency* and *fairness*, and the concept as a whole is set against *the state* and *the collective*. *Individuality*, then, like *economic freedom*, is presented as entirely positive; moreover the worth of *individualism* is assumed rather than stated: there is an assumption that the people of Britain and Hong Kong share the same values. Through this and other methods, Flowerdew argues, Patten builds up the myth of the 'good colonial'.

A similar example from Fairclough (2000) is the use of the word *business* in a corpus of British Prime Minister Blair's speeches and other texts concerned with New Labour.[3] In the New Labour corpus, *business* clearly has a positive prosody, collocating with words

[3] The Labour Party was the main socialist party in British politics. In the 1990s it revised its policies and discarded much of its socialist agenda. Party leaders now refer to it as 'New Labour'.

indicating co-operation, such as *partnership*, *involvement*, *collaboration*, *dialogue* and *relationships*, as well as with words indicating support, such as *help*, *promote*, *boost*, *empower*, *enhance*, *encourage* and so on (2000: 30–31). The implicit message is that New Labour is breaking with the traditions of the past, which set the Labour Party in opposition to the values of the business world.

Fairclough makes more explicit comparison between two corpora – the New Labour corpus and a corpus of earlier Labour Party documents – with respect to a number of words and phrases. This comparison allows him to show changes in the ideology of the party through its language or, to take a critical linguistics view, to trace the intervention of language in those changes. He also uses a general corpus for comparison. Fairclough identifies words which are disproportionately frequent in the New Labour corpus (i.e. 'keywords' in Scott's 1996 sense), such as *new*, *MODERNISE*, *partnership*, *business* and *together*. These keywords serve to identify what the proponents of New Labour see as the significant differences between their party and its predecessor. New Labour is forward-looking (not backward-looking, as Old Labour was perceived to be); it is interested in co-operation between different segments of society (not concerned with conflict, as Old Labour was perceived to be); and it is concerned with the interests of business (not anti-capitalist, as Old Labour was perceived to be). The reiteration of these keywords represents an establishment of difference between Old and New Labour.

Collocations, too, indicate the gap between Old and New Labour. The word *rights* collocates with *responsibilities* and *duties* in the New Labour corpus; conversely, *responsibilities* and *duties* collocate strongly with *rights* (Fairclough 2000: 40–41). Both rights and responsibilities are expressed as belonging to individuals. In the earlier Labour corpus, *responsibilities* is found mainly in the context of the function of public authorities and other corporate bodies. Commenting on this corpus, Fairclough observes that '. . . the close partnership between 'rights' and 'responsibilities' in New Labour language is absent, and we have rather the divorcing of rights from responsibilities . . .' (2000: 41). Another example of changing use is the word *values*, which is found much more frequently in the New Labour corpus than in the earlier one (2000: 47). In the earlier corpus, *values* (except in its economic sense) collocates most strongly with *socialist*, *socialism* or *Labour*, and occurs in almost half the instances in the context of 'conflict between Labour and its political opponents'. In the New Labour corpus, *values* does not co-occur with *socialist* or with indications of conflict; instead, *values* occurs most frequently in the context of indications of change and

modernity, as in *we have applied these values to the modern world*, and in the context of shared perceptions of decency, as in *common values*, *essential values* and *traditional values*.

Piper's work (2000a, b) examines some key items such as *lifelong learning* in a corpus of government and EU documents. Her study is wide-ranging and important for its integration of corpus observation and social theory. Here I shall concentrate on her methods of interpretation of corpus data, not because her methods are different from those of other researchers, but because she explicates them so clearly. For example, in considering the collocates that the words she is studying have, she classifies those collocates into types and then draws a connection between the collocational behaviour of the word and its social significance. For example, she notes that *learning* in a general corpus precedes words indicating problems (*disabilities*, *difficulties*) and 'experiential aspects of learning' (*curve*, *methods*, *process*, *experiences*). In the specialised 'Lifelong Learning' corpus, the same word has some of the same uses, being followed by *difficulties* and by *activities*, *methods*, *materials* and *programmes*, but it also 'modifies more institutionalised concepts', such as *society*, *age* and *culture* (Piper 2000a: 12). She comments that in these phrases 'the human subject is subsumed within superordinate and all-embracing social, cultural and temporal entities'. The term 'lifelong learning' is often preceded by *of*, and this in turn often follows words reflecting newness, such as *adoption*, *challenge*, *champion*, *implementation*, as well as words which indicate thought, such as *awareness*, *concept*, *definition*, *form*, *idea*, *issue* and *vision* (2000a: 16). Thus Piper's interpretation of *lifelong learning* as a new concept which still has to be defended with argument arises from a classification and interpretation of collocates.

As well as lexis, Piper uses grammatical concepts in her arguments. For example, she notes that the frequent phrase *for lifelong learning* most commonly follows nouns indicating provision, enablement and control (such as *foundation*, *framework*, *planning*, *policy*, *resources* and *support*) (2000a: 17). This, she notes, suggests that lifelong learning is the responsibility of institutions, who must organise it, rather than of the individual who will, hopefully, do the learning. Using Halliday's terminology, 'the people who do the learning are not agents but patients, the goals and beneficiaries of processes which are as likely to be controlled by someone else as by themselves' (2000a: 17). A related study of the term *individuals* is reported in Piper (2000b). Thus, the word, its collocates and its grammatical patterning are linked to its semantic roles and to the ideology of the texts comprising the corpus.

A somewhat similar approach is taken by Morrison and Love (1996) in their analysis of letters to the editor from issues of two Zimbabwean magazines. They use word-frequency lists to identify high-frequency content words, such as *people*, *party*, *president* and *Zimbabwe*. They then note the semantic roles that each item most often takes. For example, *President* is often the subject of verbal or mental process verbs (e.g. *the President again promised . . . as the President said himself*). Morrison and Love comment that 'the writers position the President as the articulator of government policy' (1996: 62).

The word *people*, on the other hand, is used in a number of different ways which none the less combine to form a consistent picture of oppression and suffering. *People* in Morrison and Love's corpus is the subject of clauses indicating difficulty (*. . . are facing starvation, . . . travel at least 15 kms*), or of verbs indicating a negative reaction (*. . . are grumbling, . . . are disgruntled*), or of passive verbs indicating powerlessness (*. . . are still being exploited, . . . are prevented from*). It is the object of verbs indicating 'violence and abuse' (1996: 65) (*. . . suppress the people, . . . beat up people*), and of verbs indicating manipulation of the way people think (*. . . misinforming people, . . . confusing the people*). Morrison and Love see in such patterns of use evidence of the 'discourse of disillusionment' referred to in their title.

Stubbs and Gerbig (Gerbig 1993; Stubbs 1996; Stubbs and Gerbig 1993) also focus on grammar, and on the importance of grammatical choice. Ergative verbs, such as *CLOSE*, for example, can be used to reveal or hide responsibility for a given event, as in these contrasting versions:

several firms have closed their factories
factories have been closed
factories have closed

A writer who consistently chooses the intransitive option in examples such as these presents economic events as if they were natural events, outside human control. A writer who consistently chooses the transitive, active option tends to stress the responsibility borne by people who take decisions to do things like close factories. The ideology of the text produced cannot be interpreted from a single instance of use: it is the cumulative effect of many choices that is important.

It is important also to recognise that a single grammatical choice does not have a single meaning. For example, in the discussion above, it was stated that the intransitive choice with an ergative verb

'means' that the writer or speaker fails to ascribe cause or responsibility. The implication was that the writer or speaker would be on the side of major institutions such as government, industry etc who would be the prime movers behind events such as factory closures and who may prefer to leave this unsaid. However, this is only one interpretation, Gerbig (1993, in Stubbs 1996: 145–146) argues that in texts produced by environmental agencies the intransitive in clauses such as *the crisis deepens . . . the size of the Antarctic ozone hole has increased* implies a situation that is out of control, rather than an absence of blame. Caution needs to be exercised in the interpretation of corpus data with respect to the ideology of the text-producer. This point will be returned to below.

Another example of grammatical features being used to interpret ideology is Wickens' analysis of on-line teaching materials designed for use in British university Law departments (Wickens 1998). Wickens analyses projected clauses, i.e. that-clauses following verbs such as *SAY, THINK, ARGUE, CLAIM* and adjectives such as *likely, possible, doubtful* and so on (cf Hunston 1993d; Stubbs 1996). He concentrates particularly on instances where an assertion is attributed to a source, either the speaker themselves (as in *I think the Law doesn't make it their policy to . . .*) or someone else (as in *Economists recognise that . . .*) and undertakes a careful classification of the possible sources. Wickens (1998) notes that in textbooks attributions are used to make statements about generalised circumstances and to quote other academic and legal texts. He suggests that:

> . . . the textbook is concerned to fit its knowledge claims into the linguistic structure of the academic discourse community and explicitly draws on intertextual resources to place itself within the framework of the cumulative knowledge of the field.

In seminars and lectures, on the other hand, attribution is mainly to the self, as the tutor spends some time giving his or her own opinion based on experience of the law. Quoting examples such as *Now I'm sure that the parties didn't mean that when they drafted . . .*, *And I think that that's a borderline case*, and *I was relieved to see that the House of Lords overturned this*, Wickens comments:

> [The tutor] presents the Law not as a clear cut set of rules or principles but as a fallible process which one should not take at face value.

The on-line practice materials are like the seminars in that students are presented with actual cases and are asked to discuss them in the light of legal theory and precedent. Unlike seminars, however, the on-line materials contain few statements attributed to personal opinion, of the type quoted above.

Wickens' interpretation of this is that the on-line materials are impoverished in comparison with both the textbooks and the lectures and seminars, in that students receive a less critical interpretation of what 'Law' is about. In the textbooks, Law is presented as a multi-layered, intertextual construct that is in a continual process of renegotiation. In the seminars, Law is presented as a site of personal interpretation and argument. In the on-line materials neither process – construction or interpretation – is revealed in the practices of attribution. This conclusion has implications, not just for one set of on-line materials, but for the movement towards computer-assisted learning support materials in many academic areas.

Ideology in a general corpus

In the work described above, the discourse of a particular community, in a particular context, is examined to reveal ideological implications. Writers such as Piper often use comparison with a general corpus to highlight these. Piper (2000a) notes that *learning* in her specialised corpus has collocates not found in a general corpus. On the other hand, in Piper (2000b) she notes that *individuals* has a wider range of usage in a general corpus than in the specialised one. She argues (Piper 2000b: 24) that the difference suggests that 'policy-making discourse' does not simply arise from socio-cultural norms, but quite specifically contributes to them.

Another step is to use a general corpus, not as 'background' to a specific study, but as the focus of a study itself. To do so implies regarding the general corpus as a repository of cultural information about a society as a whole (Hunston 1995b). Like the specific studies, work of this kind uses comparative frequency, collocation and phraseology, and evidence for semantic prosody as its data. Lexical items are selected for study that have a cultural saliency and which can be said to embody key ideas in a given society.

Much of the discussion below will focus on Stubbs' (1996) examples of 'cultural keywords', that is, words which capture important social and political facts about a community. Stubbs summarises his work thus:

> The main concept is that words occur in characteristic collocations, which show the associations and connotations they have, and therefore the assumptions which they embody. (Stubbs 1996: 172)

Stubbs gives numerous examples, most notably words connected

with education (*falling standards*, *back to basics* and so on), work (*worker*, *job*, *career*, *employment* and so on), nationality, and a network of interconnected words: *service*, *care*, *family* and so on. Some of the most important points he makes are these:

- Frequency information can be used to deduce what aspect of a situation the society considers to be most salient. Using simple comparative frequency, for example, Stubbs notes that in the BBC corpus from the Bank of English, the abstract word *unemployment* is much more frequent that the more personal word *unemployed*. According to Stubbs, *unemployment* 'applies to areas and populations, rather than to the people who are unemployed. It collocates not with references to individual people, but with references to groups and categories of people and to areas, and with quantitative expressions' (1996: 180). An interpretation of this is that public discourse in Britain focuses on the abstract demographic phenomenon of unemployment more frequently than on the personal experience of people who are unemployed and, by extension, that it is concerned with society as an abstract, quantifiable notion more than as a collection of individuals.
- Newly emerging collocations can be used to indicate the growth of new concepts, and changes in the meaning of words. For example, Stubbs notes that collocations such as *single parent families* and *unmarried mothers* 'signal important changes in social structures' (1996: 184). In these cases, a new phrase indicates the increasing prevalence of a particular family structure.[4] Another novel collocation, *working mother*, which means 'a mother in paid employment outside the home', is evidence for a change in the meaning of the word *work*, from a general 'doing something' to 'paid employment'.
- The range of collocates that a word has can reveal the range of associations that it has. Taking Stubbs' work on *family* a stage further, for example, it is interesting to examine the compounds consisting of *family* followed by a noun in the Bank of English. From this evidence, views of the family appear to include: the family as a cohesive group of people acting as a single entity (*family home*, *family friend*, *family holiday*, *family business*, *family reunion*, *family income*, *family gatherings*, *family outings*,

[4] In the Bank of English corpus, the phrase *one parent family/ies* occurs 319 times, *single parent family/ies* 280 times, *two parent family/ies* 111 times, and *lone parent family/ies* 29 times. The relative frequency of *one/single/lone parent family/ies* suggests the social salience of such families, whereas the significant presence of *two parent family/ies* suggests that family 'norms' are a site of conflict.

family bereavement); as a historical entity comprising continuity between generations (*family history*, *family tree*, *family tradition*, *family heirloom*); as the site of conflict and breakdown (*family therapy*, *family support*, *family breakdown*, *family problems*, *family feuds*, *family squabbles*, *family counsellor*); as a unit within the political and bureaucratic life of the state (*family law*, *family doctor*, *family health service*, *family policy*, *family unit*, *family structure*, *family allowance*); and as the site of particular social virtues (*family values*, *family entertainment*).
- Strong collocations become fixed phrases that represent a packaging of information, such that the assertion behind the phrase is less open to question than it would be in a less fixed expression. For example, Stubbs suggests that, in the context of discourse about education, the collocation *falling standards* has become a fixed phrase. It is therefore less easy to challenge the assertion that 'standards are less high now than they were previously'. He says '. . . if collocations and fixed phrases are repeatedly used as unanalysed units in media discussion and elsewhere, then it is very plausible that people will come to think about things in such terms' (1996: 195). Another example might be the phrase *illegal immigrant*. The collocation between *illegal* and *immigrant* (which has both a high t-score and a high MI-score) suggests that this is a fixed, 'unanalysed' phrase. The existence of such a fixed phrase might be said to lead people to accept without question that the movement from one country to another under some circumstances is reprehensible, and, further, that all immigration is illegitimate. (At the time of writing, the term *bogus asylum seeker* seems to be, depressingly, well on the way to becoming another such fixed phrase.) See also Dodd (2000) for examples of 'information packages' in compound nouns used in East and West German newspapers.
- Because of semantic prosody, a word or phrase can carry a covert message. For example, Stubbs (1996: 188) notes that the word *intellectual* co-occurs with words which many people would regard as negative, including *contempt*, *hippie*, *leftist* and *students*. He goes on to suggest that this negativity carries over to other collocations, such as *Jewish intellectual* and *Marxist intellectual*. He argues that, although a speaker might argue that a phrase such as *Jewish intellectual* is a purely objective description, a negative judgement is being implied, because of semantic prosody. This is perhaps controversial where a word has more than one meaning – does the prosody of one meaning carry over to the other? For example, words such as *blind* and *deaf* have 'literal' meanings

('cannot see/hear' and 'without the full range of sight/hearing') and 'metaphoric' ones. The metaphoric meanings occur in phrases such as *turn a blind eye to* and *turn a deaf ear to*. These phrases mean 'do not pay attention to', and construe the blindness and deafness in question as a deliberate avoidance strategy. It could be argued (e.g. Hunston 1999a) that the meaning of *blind* and *deaf* in these phrases constitutes a prosody that influences attitudes to literal blindness and deafness; however, there is no evidence for this influence, and a counter-argument would be that the different meanings exist independently, having no influence upon each other.

There are, it seems, two important questions in the application of corpus techniques to the study of ideology and to the practice of critical linguistics. These are: What is to be observed? and How are interpretations to be made? Stubbs, and the researchers who have been influenced by him, appear to utilise the following steps. In terms of method, frequency of occurrence, regularities of co-occurrence and of usage are observed, comparatively where appropriate. This information is used to draw conclusions about collocation, semantic prosody, and typical grammatical and semantic roles. That information in turn is used in the identification of salient concepts, of inconsistencies and sites of conflict, of difference and of change. A further level of interpretation is needed to relate these aspects to covert attitudes, implicit messages, and the discontinuity between discourse and experience.

Corpus evidence for disadvantage

One of the key concerns of critical linguistics is the perpetuation of inequality and disadvantage by a community's public discourse. There are many studies showing how groups identified by ethnicity, gender, or class, are construed in ways that constitute oppression, either in the media as a whole (e.g. van Dijk 1991, van Leeuwen 1996) or in texts of a particular type (e.g. Caldas-Coulthard 1996). More recently there have been attempts to show how information from a corpus can contribute to studies of this kind. Krishnamurthy (1996) considers the typical contexts of the words *tribe/tribal*, *ethnic/ethnicity* and *race/racial* and suggests that these words are used to construct 'otherness', in that they mark a clear difference between the groups referred to and the target readership of the texts they are used in. The use, particularly of *tribal*, is often pejorative, and Krishnamurthy argues that the use of any of these words may be evidence of racism. Similarly, Stubbs' (1996) discussion of the prosody of

intellectual (see above) suggests that the term *Jewish intellectual* is covertly anti-semitic.

Similarly, Caldas-Coulthard and Moon (1999) investigate a corpus of the *Sun* and *News of the World* newspapers to discover the adjectives collocating with words such as *man* and *woman*. The two sets of adjectives are significantly different, with only *woman* being modified significantly often by adjectives indicating physical appearance, such as *beautiful*, *pretty* and *lovely*, and only *man* being modified significantly by adjectives indicating importance, such as *key*, *big*, *great* and *main*. Further investigation confirms this asymmetry. For example, the adjective *right* is used to modify both *woman* and *man*, but *right woman* most often means 'the right woman for this man' whereas *right man* most often means 'the right man for the job' (Hunston 1999a). These findings suggest that women and men are construed differently by the discourse of popular newspapers, affirming inequality between the genders in society.

Corpus work of this kind seems to me to raise two important questions. The first is the nature of the outcome of the corpus research. One interpretation of Caldas-Coulthard and Moon's findings would run along these lines: 'We know that our society is sexist, and we know that this sexism is reflected and perpetuated by the discourse of popular journalism. The study of adjectives reveals one of the ways that these newspapers represent women and men differently and unequally.' In other words, sexism is assumed; how it is manifested is discovered. A second interpretation would be: 'Our study reveals a sexism in popular newspapers, and therefore in society, which we suspected but did not know before.' In other words, sexism is discovered; nothing is assumed. The second of these interpretations is attractive because it seems more 'objective', but it is a more difficult one to sustain. For example, I have assumed that the predominance of the personal relationship meaning of *the right woman* in a corpus of newspapers meant that women in those newspapers were construed as belonging to the domestic sphere more than to the sphere of work. The work-related meaning of *the right man* indicated that the opposite was true for men; taken together this was an argument for the perception of women as less significant in the world of paid work than men. This assumption has been challenged, however, and an alternative interpretation offered: that men are being construed as less emotionally competent because they more frequently need 'the right woman' to make their lives complete.[5] The point here is that if it is assumed, because of other evidence, that women are in

[5] This was suggested by Chris Tribble (personal communication).