1 *Introduction to a corpus in use*

What this book is about

It is no exaggeration to say that corpora, and the study of corpora, have revolutionised the study of language, and of the applications of language, over the last few decades. The improved accessibility of computers has changed corpus study from a subject for specialists only to something that is open to all. The aim of this book is to introduce students of applied linguistics to corpus investigation. Its topic is, for the most part, studies that have been carried out on corpora in English, and much of the focus of the book relates to corpora used in English language teaching. Other applications, however, such as translation and investigations of ideology, are also included. Unfortunately, the large amount of work that has been carried out on languages other than English is not covered by this book.

Although the book deals with a range of issues, there are two themes that run consistently through it. One is the effect of corpus studies upon theories of language and how languages should be described. Corpora allow researchers not only to count categories in traditional approaches to language but also to observe categories and phenomena that have not been noticed before. The other major theme is a critical approach to the methods used in investigating corpora, and a comparison between them. Corpus findings can be seductive, and it is important to be aware of the possible pitfalls in their production.

This book is intended for people who are interested in how language, more specifically English, works, and how a knowledge about language can be applied in certain real-life contexts. It is expected that the reader will wish to carry out corpus investigations for him or herself and will need to become acquainted with the range of research that has been carried out in the field.

After this introductory chapter, chapter 2 introduces some issues around corpus design and purpose, chapters 3 and 4 describe the methods used to investigate corpora, and introduce the main concepts about language that will be used in the rest of the book. Chapter 5 describes the various applications of corpora other than language teaching. Chapters 6, 7 and 8 deal with English language

teaching, chapter 6 considering new, corpus-based views of language that are relevant to teachers and chapters 7 and 8 describing some of the ways that corpora are currently influencing trends in language teaching and learning. Chapter 9 concludes the book.

Before continuing, it is worth asking two questions about the title of this book: what is a corpus? and what is applied linguistics?

A corpus is defined in terms of both its form and its purpose. Linguists have always used the word *corpus* to describe a collection of naturally occurring examples of language, consisting of anything from a few sentences to a set of written texts or tape recordings, which have been collected for linguistic study. More recently, the word has been reserved for collections of texts (or parts of text) that are stored and accessed electronically. Because computers can hold and process large amounts of information, electronic corpora are usually larger than the small, paper-based collections previously used to study aspects of language. A corpus is planned, though chance may play a part in the text collection, and it is designed for some linguistic purpose. The specific purpose of the design determines the selection of texts, and the aim is other than to preserve the texts themselves because they have intrinsic value. This differentiates a corpus from a library or an electronic archive. The corpus is stored in such a way that it can be studied non-linearly, and both quantitatively and qualitatively. The purpose is not simply to access the texts in order to read them, which again distinguishes the corpus from the library and the archive.

The field of applied linguistics itself has undergone something of a revolution over the last few decades. Once, it was almost synonymous with language teaching but now it covers any application of language to the solution of real-life problems. As has often been said (e.g.Widdowson 1979; 2000), the difference between linguistics and applied linguistics is not simply that one deals with theory and the other with applications of those theories. Rather, applied linguistics has tended to develop language theories of its own, ones that are more relevant to the questions applied linguistics seeks to answer than are those developed by theoretical linguistics. Increasingly, corpora are adding to the development of those applied views of language.

The rest of this chapter will give an overview of what a corpus can do and how corpora are used in applied linguistics. This is followed by an account of the main types of corpora and an introduction to some of the terminology used in this book. The chapter concludes with a discussion of the advantages and limitations of using corpora in language study.

What a corpus can do

Strictly speaking, a corpus by itself can do nothing at all, being nothing other than a store of used language. Corpus access software, however, can re-arrange that store so that observations of various kinds can be made. If a corpus represents, very roughly and partially, a speaker's experience of language, the access software re-orders that experience so that it can be examined in ways that are usually impossible. A corpus does not contain new information about language, but the software offers us a new perspective on the familiar. Most readily available software packages process data from a corpus in three ways: showing frequency, phraseology, and collocation. Each of these will be exemplified in this section.

Frequency

The words in a corpus can be arranged in order of their frequency in that corpus. This is most interesting when corpora are compared in terms of their frequency lists. Table 1.1 shows the top 50 words in a corpus of politics dissertations compared with a comparable corpus of materials science dissertations (data from Charles, in preparation) and with the 1998 Bank of English corpus (data from Sinclair 1999).

In all three corpora, grammar words are more frequent than lexical words; indeed, the words *the, of, to, and, a* and *in* occupy the top six places in each corpus. The only lexical word which comes into the top 50 words of the general Bank of English corpus is *said* (at number 36). The lexical words in the other corpora reflect their subject matter, e.g. *surface* (34), *energy* (37), *electron* (48) and *particles* (50) in materials science, and *international* (21), *policy* (28), *states* (29) and *socialization* (50) in politics. There are more such words in the materials science list than in the politics list. One reason for this might be that in materials science the prose is more dense, with more lexical words occurring together without grammar words between them (cf Halliday and Martin 1993: 76–77), as in *electron probe microanalyser, electron spin resonance dating techniques* and *high electron mobility transistor.* Another reason might be that in materials science the vocabulary is less wide than in politics, so that fewer words appear more frequently. One of the notable features of the grammar words in the lists is that *this* occurs much higher up the materials science list (8) and the politics list (13) than the general corpus list (28). *This* is often used to summarise what has been said before, as in *'mind' and 'mental' processes are now respectable concepts in psychology . . . This is important not only*

Table 1.1. Word frequency comparisons across corpora

	General corpus	Materials science	Politics
1	THE	THE	THE
2	OF	OF	OF
3	TO	AND	TO
4	AND	IN	AND
5	A	TO	IN
6	IN	A	A
7	THAT	IS	THAT
8	S	THIS	IS
9	IS	P	AS
10	IT	THAT	WAS
11	FOR	FOR	FOR
12	I	BE	IT
13	WAS	AS	THIS
14	ON	HEAD	P
15	HE	ARE	ON
16	WITH	WITH	BE
17	AS	IT	BY
18	YOU	BY	WHICH
19	BE	ON	S
20	AT	WAS	NOT
21	BY	AT	INTERNATIONAL
22	BUT	WHICH	WITH
23	HAVE	FROM	AN
24	ARE	FIGURE	QUOTE
25	HIS	AN	ARE
26	FROM	NOT	FROM
27	THEY	HAS	WERE
28	THIS	WERE	POLICY
29	NOT	CAN	STATES
30	HAD	THESE	BUT
31	HAS	BEEN	STATE
32	AN	HAVE	WOULD
33	WE	OR	OR
34	N'T	SURFACE	ITS
35	OR	USED	MAZZINI
36	SAID	C	THEIR
37	ONE	ENERGY	HEAD
38	THERE	TEMPERATURE	AT
39	WILL	ALSO	HAD
40	THEIR	WILL	HAVE
41	WHICH	CONTRAST	MORE
42	SHE	TWO	BRITAIN
43	WERE	FIELD	THEY

Table 1.1 (*cont.*)

	General corpus	Materials science	Politics
44	ALL	SAMPLE	THESE
45	BEEN	MATERIAL	HE
46	WHO	CURRENT	BETWEEN
47	HER	BETWEEN	HIS
48	WOULD	ELECTRON	US
49	UP	HOWEVER	THAN
50	IF	PARTICLES	SOCIALIZATION

Note: In the corpora from which this table is derived, 'C' and 'P' are symbols and abbreviations, such as the abbreviation for *centigrade*. 'P' is sometimes also the code marking a new paragraph. 'S' is usually the 's' following an apostrophe, as in John'*s* or she'*s*.

for psychologists, but for society in general, where *This* summarises the preceding sentence. However, this use is more common in written argument, and therefore in academic prose, than in speech or in writing that is more speech-like. Other words associated more with speech and informal writing than with formal writing, such as *I, but* and *n't* occur in the general corpus list but not in the other two.

Frequency lists from corpora can be useful for identifying possible differences between the corpora that can then be studied in more detail. Another approach is to look at the frequency of given words, compared across corpora. Table 1.2 shows the number of occurrences of *must, have to, incredibly* and *surprisingly* in three corpora from the Bank of English: a corpus of books published in Britain, a corpus of *The Times* newspaper, and a corpus of spoken British English. Because the three corpora are of different sizes, a comparison of actual frequencies would not be useful, so the figures for occurrences per million words are given. For example, the *Times* corpus is nearly 21 million words. There are about 9,600 occurrences of the word *must* in that corpus, giving a frequency per million of just over 460.

Table 1.2 can be used to compare *must* with *have to*, and *incredibly* with *surprisingly*. Whereas the books corpus and the *Times* corpus use *must* in preference to *have to*, the spoken corpus shows the reverse trend, suggesting that *have to* is less formal than *must*. Similarly, *surprisingly* is found less frequently in the spoken corpus than in the other two, whilst for *incredibly* the reverse is true. This suggests that *incredibly* is a less formal word than *surprisingly*. Whilst this appeal to 'formality' may offer partial insight, a more satisfactory explanation can be found by looking at the words more closely.

Table 1.2. Frequencies of *must*, *have to*, *incredibly* and *surprisingly* across corpora (per million words)

	Books	*Times*	Spoken
must	683	460	363
have to	419	371	802
Total	1102	831	1165
incredibly	8	10	15
surprisingly	25	29	4
Total	33	39	19

In the three corpora mentioned, *incredibly* is used almost exclusively before an adjective or adverb, the most significant being *difficult*, *well*, *important*, *hard*, *complex* and *strong*. Here are some examples of typical uses:

Well I mean now as I'm unemployed with fairly specialist skills erm I find it incredibly difficult to find work that is suitable. (spoken corpus)

Why on earth was she standing here blubbing like a baby at her age? She should be proud at this moment. Noora had done incredibly well to get this far in so short a time. (books corpus)

But I was fascinated by it all to find out how this incredibly important woman operates, what she's really like, how she thinks, the whole upstairs-downstairs thing. (*Times* corpus)

The word *surprisingly* shares some of this behaviour: the words *good*, *little*, *large*, *few*, *well* and *strong* appear significantly frequently after it in the three corpora mentioned, as exemplified here:

The reason motorcycles have become popular inner-city transport owes much to the machines and the protective clothing now on the market. The machines are powerful, stylish and comfortable, and their aerodynamics give surprisingly good weather protection. (*Times* corpus)

As a society we are ill-informed about epilepsy, often finding it shocking and something we would prefer not to be exposed to rather than an illness. The sufferers often receive surprisingly little support either within their family or from colleagues, employers or friends. (books corpus)

I'm going to write some recommendation as to . . . how to publicize it if people don't know about it . . . Erm and so far erm surprisingly few people know about it. Erm I'd have thought more would know but they don't. (spoken corpus)

Looking at these examples it seems that *surprisingly* is used to mean 'contrary to expectation' whereas *incredibly* is used as a strong version of 'very'. This goes some way to explaining why *incredibly* is more frequent in spoken English than in written. The adverb *surprisingly* also has a use which *incredibly* does not have. As well as being followed by an adjective or adverb, it is also followed significantly often by a word that is the beginning of a clause, such as *he, the* or *it*. It is also often preceded by *not*, *perhaps* or *hardly*. This indicates that *surprisingly* is used to modify a clause as well as to modify an adjective or adverb, as in these examples:

Woan, now 28 . . . was rejected in his teens by Everton, indulged in non-league football with Runcorn until he was 22, and studied by day to be a chartered surveyor. Not surprisingly, he reads books more than most footballers do, and his recent favourite was *Extraordinary Power* by Joseph Finder. (*Times* corpus)

There was another shop just around the corner where I had to catch a second bus to Clapton and there, hardly surprisingly, the news that Sandown had indeed fallen victim to the elements was received with much regret. (books corpus)

Now having said that you then have the opposite problem that by being exhaustive everybody goes to sleep or throws the thing in the bin erm and not surprisingly it has been found that the first two or three items are attended to considerably more than the three hundred and thirtieth. (spoken corpus: from a seminar on survey techniques)

Although this use of an adverb to modify a clause does occur in some registers of spoken English, as the last example above shows, it is a feature not associated with colloquial speech. This adds another reason for the difference in frequency among the corpora.

Another example of differences in frequency is the words *man*, *woman*, *husband* and *wife*. Table 1.3 shows the frequencies (i.e. the number of occurrences per million words) in the same three corpora, and the total frequency across those three corpora.

The totals show that *man* occurs more frequently than *woman*, and it is therefore unexpected that *wife* should occur more frequently than *husband*. The most likely interpretation is that women are relatively more frequently referred to in relation to the person they are married to than men are. This seems to be confirmed by a more detailed investigation of the *Times* corpus, in which the phrase *husband of* occurs 53 times (2.5 times per million words) whereas the corresponding *wife of* occurs 299 times (14.3 times per million words). A typical instance is the description of a woman as *a top US model and wife of Gregory Peck's son*, illustrating (twice!) how

Table 1.3. Frequencies of *man*, *woman*, *husband* and *wife* across corpora (per million words)

	Books	*Times*	Spoken	Total
man	980	583	285	1848
woman	456	208	137	801
husband	163	140	92	395
wife	216	224	83	523

less famous people tend to be described in terms of their more famous relatives. Although the equivalent *husband of* is used in those cases where the wife is more famous, the frequency figures indicate that it is less usual for a woman to be more noteworthy than her husband.

The spoken corpus, however, reverses the trend apparent in the books and *Times* corpora. In that corpus, *husband* is more frequent than *wife*, just as *man* is more frequent than *woman*. In both cases the most frequent phrases are with possessive determiners: *my husband, his wife* and so on. Although *wife of* is fairly frequent (28 instances) and much more frequent than *husband of* (only 8 instances), this form of the possessive is less significant than in the *Times* corpus. In the *Times*, 6% of the instances of *wife* comprise the phrase *wife of*, whereas in the spoken corpus the figure is 1.6%. It seems, then, that one explanation for the discrepancy in figures between *husband* and *wife* is accounted for by the tendency to relate 'unknown' people to 'known' ones, a tendency which occurs in some registers much more than in others. This tendency in published written discourse might be argued to perpetuate discrimination against women.

More sophisticated work on comparative frequencies between registers has been undertaken by Biber and his colleagues (e.g. Biber 1988; Biber et al 1998; Biber et al 1999; see also Mindt 2000 and Leech et al 2001). They use software which counts not only words but also categories of linguistic item. One example among many is their calculation of the distribution of present and past tenses across four registers: 'conversation', 'fiction', 'news' and 'academic' (Biber et al 1999: 456). They note that in their conversation and academic corpora, present tense occurs more frequently than past tense. In the fiction corpus, the opposite is the case, with past tense preferred to present tense. In the news corpus, the figures are roughly equal. These findings may be seen in the context of Halliday's (1993) calculation that in the Bank of English, present and past tenses are

found in roughly equal proportions. Common sense suggests that it is reasonable to extrapolate from these findings a statement about English as a whole. Each register has its own ratio of present and past, but overall the figures balance out, and a 50:50 proportion is maintained. However, Biber et al's findings also sound a warning in interpreting Halliday's figures. If the proportion of present to past is dependent on register, then the proportion in a large corpus will in turn depend on the balance of registers within that corpus. Too much fiction, for example, will bias the figures towards past tense. As will be discussed in chapter 2, however, this is a far from simple matter to resolve. How much fiction would be 'too much'? As we have no idea how to calculate proportions for 'English as a whole', we have equally no idea what would constitute a corpus that truly reflected English.

Phraseology

Most people access a corpus through a concordancing program. Concordance lines bring together many instances of use of a word or phrase, allowing the user to observe regularities in use that tend to remain unobserved when the same words or phrases are met in their normal contexts. (Sinclair and Coulthard 1975 used the term *latent patterning* to refer to this phenomenon.) It is through concordances, then, that phraseology is observed.

As much of chapter 3 of this book will consider phraseology in some detail, it will be dealt with only briefly here. One point of interest is the way that phraseology can be used as an alternative view of phenomena that teachers of English are frequently called upon to explain. For example, learners often confuse adjectives such as *interested* and *interesting*, and find that explanations of the different meanings do not make the choice more accessible in spontaneous speech. Below are 23 lines each of the words *interested* and *interesting* (selected at random from the Bank of English).

Interested

and the surrounding areas who are interested in water sports. Rural
than Barbados. If you are interested in wild-life, Tobago is heaven,
The YOA claims to be interested in lobbying on issues, but it
irony in that, whereas I'm more interested in the musical arrangements ad
like bigger speakers. I'm more interested in playing videos. I've got a
work or whatever that you might be interested in speakers.
the new test. MORTIMER: We've been interested in looking at alternative methods
by around half a dozen firms interested in acquiring its Welsh business.
over a Labour Party which was less interested in evangelism than it was in the
on radon, but they tend to be more interested in measuring it once it escapes

and from the outside and – I was interested in something somebody said about
(another ambiguity), became interested in African cash-crops, that it
to say, you're going to say she's interested in my money. I expected a
all his readings, he appeared more interested in developing independent grown-
those maps their due, it is more interested in reconstructing the maps etched
users and the medical company interested in the product. Even when venture
and what then?" Yes, he is interested in moving towards contemporary
that the Woodland Trust charity is interested in maintaining woodlands and we
in case of emergency and may be interested in a car kit for hands-free
on the work of Henry Miller first interested me in the subject. And now I am
know?' I do. Not to interfere. I'm interested.' OK. Do you want to come now?'
make decisions. Insurers are interested too; they want to use such
consultants have been interested you know in various systems. And

Interesting

Yeah. Yeah. But there's this interesting annual variation and erm I
rights legislation. Now this is interesting, Bob. Here's one that you would
trusted her." In one of the most interesting chapters of this biography,
Sunshine Sprint runners, but one interesting entry yesterday was the Glenn
Heal. Well, it's been an interesting few days for the Liberal
game." He might have added an interesting historical fact: The last Series
or in auto accidents. It's a very interesting idea because that could largely
image of the object. What is interesting is that it is not necessary to
ll mature into something big and interesting like REM, won't the music
conservation of some biologically interesting niches could also be brought
yard bet on Foyt, just to make it interesting. 'Not Andretti?' Tucker asked.
I learned a lot about geology, met interesting people and went home with a good
antiques, appropriate fabrics and interesting pictures. (There are no bar-
249), but this was hardly the most interesting point in his view. The
to be Master of Wine, argue an interesting proposition: 'At dollar 180,
the Republicans. It should be an interesting ride. When liberals were
just as things have started to get interesting, the film comes to an abrupt
curse which says, 'May you live in interesting times.' Well, here I am – living
e the game away but erm phoo er be interesting to see how people fall on that
Unidentified Man 1: It's interesting to – to know, but it doesn't
and style; it would be interesting to see what this McAllister pup
that has started to produce really interesting wines, especially whites, at &
Oh yes. That's very interesting yeah. If he didn't get

What these lines show is that, overwhelmingly, *interested* is used in the phrase *interested in*, and the pattern 'someone is interested in something' is exceptionally frequent. By contrast, *interesting* is nearly always used before a noun, in the pattern 'an interesting thing'. Significant exceptions to this include 'What is interesting is . . .' and 'It's interesting to see . . .' The minimal pair that might be represented by 'the boy is interested' and 'the boy is interesting' occurs comparatively rarely (though it must be remembered that 23 lines are only a small proportion of the total). The focus of what is to be taught, therefore, shifts from the confusable pair *interested* and *interesting* to the phrases 'someone is interested in something', 'an

interesting thing', 'what is interesting is' and 'it is interesting to see', which are, hopefully, different enough to be less easily confused.

A similar approach is taken by Kennedy (1991) in his study of *between* and *through*. After pointing out that reference books have difficulty in expressing the differences between these words, Kennedy adopts a phraseological approach. He notes that *between* is frequently found after nouns such as *difference, distinction, gap, contrast, conflict* and *quarrel*, as well as *relationship, agreement, comparison, meeting, contact* and *correlation*, whereas *through* is more frequently found after verbs such as *go, pass, come, run, fall* and *lead*. These and other observations enable him to provide a profile of each word (1991: 106–107) that relates each aspect of meaning to typical phraseologies. Kennedy is also able to assign frequencies to the different meanings, or 'semantic functions'. Approximately a quarter of the instances of *between* in the Lancaster-Oslo-Bergen (LOB) corpus have a 'location' meaning (e.g. *the channel between Africa and Sicily; earnings between £5 and £6 a week*) whilst about the same proportion of the instances of *through* have an 'instrumental' meaning (e.g. *I should have met him through Robert Graves; evidence obtained through the examination of stones*).

Phraseology of this kind can be an extremely subtle phenomenon. Below are all the instances of the phrase *GRASP the point* from the Bank of English, with the lines numbered for reference. (Note: here and henceforward, capitals are used to indicate all the forms of a verb. For example, *GRASP* means *grasp*, *grasps*, *grasping* and *grasped*.)

1 in Free, where Teenotchy tries to grasp the point and the structure of
2 accident will help the islanders grasp the point. Racing: Guardian's
3 Please, all of you out there, try to grasp the point. We do not want
4 the Independent on Sunday. I fail to grasp the point of newspapers' divided
5 wonderful. People are able at last to grasp the point of it now that the
6 some scholars were beginning to grasp the point, most shared the
7 Beginning. When you want readers to grasp the point of a paragraph right
8 is likely to be his failure to grasp the point made by his former pri
9 incompetent, often envious, rarely grasp the point of any given book, if
10 always supposing the latter has yet grasped the point – and has responded
11 of it now. He doesn't seem to have grasped the point of the project.
12 me, I'd kill him." But when they had grasped the point of it, they became
13 this hornet's nest. Giovanni Benelli grasped the point at once. He saw Worl
14 genes in the cell. Once we have grasped the point about genes working
15 members – if they hadn't already grasped the point – that 'money has be
16 Belatedly, Yasser Arafat has grasped the point that his people in
17 team do not seem to have grasped the point about these jokes:
18 if not all communication that one grasps the point of what someone is

A simple observation here is that *point* is frequently followed by *of*