Corpora in Applied Linguistics

Susan Hunston
University of Birmingham

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, UK

www.cambridge.org
Information on this title: www.cambridge.org/9780521801713

First published 2002
3rd printing 2005

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

ISBN-13 978-0-521-80171-3 hardback
ISBN-10 0-521-80171-0 hardback

ISBN-13 978-0-521-80583-4 paperback
ISBN-10 0-521-80583-X paperback

Contents

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*

(lines 4, 5, 7, 9, 11, 12, 18). Less obvious, perhaps, is what comes before *GRASP*. In most cases, there is either an indication of something not being done (lines 4, 8, 9, 11, 17) or an indication of something difficult being achieved (lines 1, 2, 3, 5, 6, 7, 10, 15, 16). Even when a line appears to be a counter-example, as in line 13, a look at more co-text indicates that 'grasping the point' is problematic even here: in this case, Benelli is contrasted with another person who fails to grasp the point. This subtlety of usage is difficult to intuit, and is observable only when a lot of evidence is seen together so that the pattern emerges.

Collocation

The final example of how the data in corpora can be manipulated is the calculation of collocation. This will be examined in more detail in chapter 4; here it is sufficient to note that collocation is the statistical tendency of words to co-occur. A list of the collocates of a given word can yield similar information to that provided by concordance lines, with the difference that more information can be processed more accurately by the statistical operations of the computer than can be dealt with by the human observer.

For example, collocates of the word *shed* include: *light*, *tear/s*, *garden*, *jobs*, *blood*, *cents*, *image*, *pounds*, *staff*, *skin* and *clothes*. Only when it collocates with *garden* is the word *shed* a noun; in all other cases it is a verb. Its meaning is something like 'lose' or 'give', but the precise meaning of each phrase depends on the collocate:

shed light (on) means 'illuminate', usually metaphorically;
shed tears means 'cry' (literally) or 'be sorrowful' (crying metaphorical tears);
shed blood means 'suffer' or 'die', either literally or metaphorically;
shed jobs and *shed staff* mean 'get rid of people';
shed pounds means 'lose weight';
in *shed skin* and *shed clothes*, *shed* means 'remove';
shed cents is used to indicate that shares or a currency become reduced in value;
shed image means a deliberate changing of how one is perceived.

Collocation can indicate pairs of lexical items, such as *shed* + *tears*, or the association between a lexical word and its frequent grammatical environment.[1] For example, the word *head* has the following lexical collocates:

[1] The collocation between a lexical word and a grammatical one is frequently termed 'colligation'.

SHAKE, *injuries* and *SHOOT*, in which *head* indicates a part of the body (as in *shook her head*; *head injuries*; *was shot in the head*, for example);
state, *office*, *former* and *department*, in which *head* indicates a person in charge;

and these grammatical collocates:

possessives such as *his*, *her*, *my* and *your*;
of, used in phrases such as *head of department*;
over, used in phrases such as *HIT/BEAT someone over the head*, *HOLD something over someone's head*, *GO over one's head* and *LOSE one's head over someone*;
on, as in *HIT someone on the head*, *PUT something on one's head*, *MEET something head on*, and *TURN something on its head*;
back, as in *back of the head*, *head back* and *PUT/THROW one's head back*;
off, as in *head off a problem*, *CUT someone's head off* and *head off towards somewhere*.

The various phraseologies of *head*, together with the meanings associated with these phraseologies, are indicated by these collocates.

What corpora are used for

Corpora nowadays have a diverse range of uses, which will be discussed more fully in chapters 5, 7 and 8, but some are summarised here:

- For language teaching, corpora can give information about how a language works that may not be accessible to native speaker intuition, such as the detailed phraseology mentioned above. In addition, the relative frequency of different features can be calculated. According to Mindt (2000), for example, nearly all the future time reference in conversational English is indicated by *will* or other modals. The phrase *BE going to* accounts for about 10% of future time reference, and the present progressive less than 5%. Information such as this is important for syllabus and materials design.
- Increasingly, language classroom teachers are encouraging students to explore corpora for themselves (see, for example, Burnard and McEnery eds. 2000), allowing them to observe nuances of usage and to make comparisons between languages.
- Translators use comparable corpora to compare the use of apparent translation equivalents in two languages, and parallel

corpora to see how words and phrases have been translated in the past. As an example in chapter 5 shows, for example, the English word *still* can translate or be translated by the French *toujours* or *encore*, or by expressions with *couramment* or the verb *continuer*. Sometimes when an English sentence includes the word *still* the parallel French sentence has no translation equivalent at all, but when *toujours* and *encore* are present in the French sentence, the English parallel sentence always contains *still*.

- General corpora can be used to establish norms of frequency and usage against which individual texts can be measured. This has applications for work in stylistics and in clinical and forensic linguistics.
- Corpora are used also to investigate cultural attitudes expressed through language (e.g. Stubbs 1996; Teubert 2000) and as a resource for critical discourse studies (e.g. Krishnamurthy 1996; Caldas-Coulthard and Moon 1999; Fairclough 2000).

Types of corpora

A corpus is always designed for a particular purpose, and the type of corpus will depend on its purpose. Here are some commonly used corpus types:

- **Specialised corpus**. A corpus of texts of a particular type, such as newspaper editorials, geography textbooks, academic articles in a particular subject, lectures, casual conversations, essays written by students etc. It aims to be representative of a given type of text. It is used to investigate a particular type of language. Researchers often collect their own specialised corpora to reflect the kind of language they want to investigate. There is no limit to the degree of specialisation involved, but the parameters are set to limit the kind of texts included. For example, a corpus might be restricted to a time frame, consisting of texts from a particular century, or to a social setting, such as conversations taking place in a bookshop, or to a given topic, such as newspaper articles dealing with the European Union. Some well-known specialised corpora include the 5 million word Cambridge and Nottingham Corpus of Discourse in English (CANCODE) (informal registers of British English) and the Michigan Corpus of Academic Spoken English (MICASE) (spoken registers in a US academic setting).
- **General corpus**. A corpus of texts of many types. It may include written or spoken language, or both, and may include texts produced in one country or many. It is unlikely to be representative

of any particular 'whole', but will include as wide a spread of texts as possible. A general corpus is usually much larger than a specialised corpus. It may be used to produce reference materials for language learning or translation, and it is often used as a baseline in comparison with more specialised corpora. Because of this second function it is also sometimes called a **reference corpus**. Well-known general corpora include the British National Corpus (100 million words) and the Bank of English (400 million words in January 2001),[2] both of which comprise a range of sub-corpora from different sources. Much earlier general corpora were the LOB corpus, consisting of written British English, and the Brown corpus, consisting of written American English, both compiled in the 1960s and comprising 1 million words each.

- **Comparable corpora**. Two (or more) corpora in different languages (e.g. English and Spanish) or in different varieties of a language (e.g. Indian English and Canadian English). They are designed along the same lines, for example they will contain the same proportions of newspaper texts, novels, casual conversation, and so on. Comparable corpora of varieties of the same language can be used to compare those varieties. Comparable corpora of different languages can be used by translators and by learners to identify differences and equivalences in each language. The ICE corpora (International Corpus of English) are comparable corpora of 1 million words each of different varieties of English.
- **Parallel corpora**. Two (or more) corpora in different languages, each containing texts that have been translated from one language into the other (e.g. a novel in English that has been translated into Spanish, and one in Spanish that has been translated into English) or texts that have been produced simultaneously in two or more languages (e.g. European Union regulations, which are published in all the official languages of the EU). They can be used by translators and by learners to find potential equivalent expressions in each language and to investigate differences between languages.
- **Learner corpus**. A collection of texts – essays, for example – produced by learners of a language. The purpose of this corpus is to identify in what respects learners differ from each other and from the language of native speakers, for which a comparable corpus of native-speaker texts is required. There are a number of learner corpora around the world, of which the best known is the

[2] The Bank of English has increased in size progressively since its inception in the 1980s. Some of the studies in this book were done on the 323 million word corpus; others on the 400 million word one.

International Corpus of Learner English (ICLE). This is in fact a collection of corpora of 20,000 words each, each one comprising essays written by learners of English from a particular language background (French, Swedish, German etc). There is a comparable corpus of essays written by native speakers of English: the Louvain Corpus of Native English Essays (LOCNESS).

- **Pedagogic corpus**. A corpus consisting of all the language a learner has been exposed to. For most learners, their pedagogic corpus does not exist in physical form. If a teacher or researcher does decide to collect a pedagogic corpus, it can consist of all the course books, readers etc a learner has used, plus any tapes etc they have heard. The term 'pedagogic corpus' is used by D. Willis (1993). A pedagogic corpus can be used to collect together for the learner all instances of a word or phrase they have come across in different contexts, for the purpose of raising awareness. It can also be compared with a corpus of naturally occurring English to check that the learner is being presented with language that is natural-sounding and useful.
- **Historical or diachronic corpus**. A corpus of texts from different periods of time. It is used to trace the development of aspects of a language over time. Perhaps the best-known historical corpus of English is the Helsinki Corpus, which consists of texts from 700 to 1700 and comprises 1.5 million words.
- **Monitor corpus**. A corpus designed to track current changes in a language. A monitor corpus is added to annually, monthly or even daily, so it rapidly increases in size. However, the proportion of text types in the corpus remains constant, so that each year (or month or day) is directly comparable with every other.

Issues in the design and purpose of corpora are discussed in chapter 2.

Some key terms

The literature on corpora makes use of a certain amount of technical terminology. It may be helpful to explain a few of the most essential terms here. Eight terms will be explained: *type*, *token*, *hapax*, *lemma*, *word-form*, *tag*, *parse* and *annotate*.

'Type', 'token' and 'hapax'

How many words are there in the following paragraph (taken from Simpson and Montgomery 1995: 140)?

What elements make up a narrative? Providing an answer to this question has become one of the central challenges for a stylistics of prose fiction. Much work in modern narrative stylistics seeks to isolate the various units which combine to form a novel or short story and to explain how these narrative units are interconnected. Having identified the basic units in this way, the next task is to specify which type of stylistic model is best suited to the study of which particular unit.

In one sense, there are 84 words. That is, there are 84 sequences of letters separated by spaces or punctuation. This is the figure that the word-count function of a word-processing program gives. In other words, there are 84 **tokens**.

However, many of these words occur more than once:

a	occurs 3 times
narrative	occurs 3 times
to	occurs 6 times
this	occurs 2 times
of	occurs 4 times
the	occurs 5 times
in	occurs 2 times
stylistics	occurs 2 times
units	occurs 3 times
which	occurs 3 times
is	occurs 2 times

Counting each repeated item once only, so that only different words are counted, gives a total of 60 items. Using the terminology, there are 60 **types**. The words that occur only once are called **hapax legomena** or **hapaxes**.

The short sample paragraph, therefore, comprises a corpus of 84 tokens and 60 types, including 33 hapaxes. In a very small corpus like this, the ratio of types to tokens might be expected to be high as in this example. In a larger corpus, there will be relatively more tokens for each type, as there is more repetition of individual words in longer texts (Biber et al 1999: 53).

'Lemma' and 'word-form'

There is a further factor to be taken into account when dealing with 'words' in a corpus. In the paragraph about narrative quoted above, for example, it could be argued that *unit* and *units* are in a sense the 'same word', in that one is simply the plural form of the other. We might say, then, that *unit* and *units* are two **word-forms** belonging to the same **lemma**: *UNIT*. In the same way, *eat*, *eats*, *eating*, *ate* and

eaten are word-forms belonging to the lemma *EAT*. There is some debate as to whether two word-forms belong to the same lemma if they belong to different word-classes (for example, do the adjective *stylistic* and the noun *stylistics* in the above paragraph belong to the same lemma or not?). To a large extent the notion of lemma is a convenience, so what is to be counted in a lemma depends on what use the idea is to be put to. Often, for example, it is useful to see what prepositions follow an adjective by getting all instances of the adjective lemma only. In this book, unless otherwise stated, word-forms will be said to belong to the same lemma only if they belong to the same word-class. Thus, *quick*, *quicker* and *quickest* belong to the lemma *QUICK*, but *quickly* will not be said to belong to the same lemma.

'Tag', 'parse' and 'annotate'

These terms refer to procedures that are carried out to add information to the words in a corpus. The additions may be made automatically (i.e. by a computer program alone) or manually (i.e. by a human being working with the computer program). Adding information automatically is a fast and easily repeated process, but often of limited accuracy; adding information manually is a relatively slow process, and needs to be repeated if the corpus is changed or enlarged, but the results are more accurate.[3] Speed and accuracy are two of the key issues in the addition of information to a corpus.

The term **tagging** is normally used to refer to the addition of a code to each word in a corpus, indicating the part of speech. It is feasible to tag a corpus automatically, and such tagging will be reasonably, but not entirely accurate. For a small corpus it is possible to edit the tags to obtain a higher degree of accuracy; for a very large corpus, this is not normally practical.

Tags are useful as components of word searches. Someone wishing to investigate a word such as *work*, for example, may wish to look at the nouns separately from the verbs. A tagger allows these to be searched for independently. In addition, a tagger may be used to make calculations of proportions of word use. For example, Granger and Rayson (1998) compare native and non-native corpora and

[3] For the purposes of this discussion, 'accurate' means 'what a competent human analyst would decide'. The issue of accuracy is not this simple, however. Human analysts make mistakes if they are tired or bored; computers do not become tired, and in that sense they are 'more accurate'. Sometimes even the notion of 'accuracy' itself is misleading. For example, both human beings and computers may argue about whether *circle* in *circle line* is a noun or an adjective.

report that the non-native speakers use more determiners, pronouns and adverbs than native speakers do, but fewer conjunctions, prepositions and nouns. A possible explanation for this is that the native speakers use more complex and abstract noun phrases than the non-native speakers do. As another example, it is possible to count the number of nouns, verbs, and so on, in a corpus as a whole, or to calculate whether a particular word is more frequently used as a noun or a verb in a given register. In the *New Scientist* corpus of the Bank of English, the verb *WORK* occurs 926 times per million words, and the noun occurs 654 times. In the spoken corpus, the verb occurs 1,060 times per million words, and noun 572 times. Although in both corpora the verb is more frequent than the noun, the noun is relatively less frequently used in the spoken corpus than in the *New Scientist* corpus. Looking at the collocations of the noun *WORK* in each case suggests differences in meaning too. The most significant collocates in the *New Scientist* are the possessives, especially in the phrases *their work* (i.e. the investigations of a group of scientists) and *the work of*. In other words, a frequent meaning of the noun *WORK* is scientific discovery. Another frequent meaning is to describe what non-human entities, such as bacteria, do. In the spoken corpus, significant collocates include *of work* (as in *loads of work, sort of work*) and *at work*. The most frequent meaning of the noun *WORK* is 'job'. The meaning of *WORK* (noun) in the *New Scientist* is more central in terms of the topic of that magazine than the meaning of the noun in the spoken corpus is, and this perhaps explains why the noun is relatively more frequent in the *New Scientist* corpus.

Corpus **parsing** is the analysis of text into constituents, such as clauses and groups. A parsed corpus can be used to count with great accuracy the number of different structures in a corpus.

Parsing can be done automatically, but the resulting output is often not very accurate. Accuracy can be improved by 'training' the automatic parser, that is, by setting up the parser to learn from past examples. In that case, a small corpus is parsed and edited manually and the resulting output is used to train the automatic parser (Leech and Eyes 1997). Parsers of this level of sophistication have been developed by Leech and his colleagues at Lancaster University and, though the process is somewhat cumbersome, a high level of accuracy is achieved. Where total accuracy is required, however, for example where the parsed text is being used to teach human learners how to do grammatical parsing, manual editing is still needed (McEnery et al 1997).

A superordinate term for tagging and parsing is **annotation.** 'Annotation' is also used to describe other kinds of information that

can be added to a corpus. Again, Lancaster University leads this field, and numerous interesting examples can be found in Garside et al (eds. 1997). Annotation in this more limited sense is often done manually. Typical examples include: the annotation of a spoken corpus for intonation; annotation for anaphora, which identifies the cohesive item and its referent; and annotation of various means of representing speech and thought in written text. Annotation of this kind essentially uses the computer to keep track of very lengthy manual analyses, so that the statistics from these analyses can more easily be calculated. (See chapter 4 for further examples of corpus annotation.)

Why corpora? Why not?

At the end of this introductory chapter, it is worth returning to the central question of why corpora are important for applied linguists, and also to consider their limitations. Corpora are often described as a tool, and the development of corpora has been likened to the invention of telescopes in the history of astronomy (Stubbs 1996: 231). It might be more proper to say that corpora are a way of collecting and storing data, and that it is the corpus access programs – presenting concordance lines and calculating frequencies – that are the tools. Stubbs (1999) points out that, just as it is ridiculous to criticise a telescope for not being a microscope, so it is pointless to criticise corpora for not allowing some methods of investigation. They are invaluable for doing what they do, and what they do not do must be done in another way.

A corpus essentially tells us what language is like, and the main argument in favour of using a corpus is that it is a more reliable guide to language use than native speaker intuition is. Although a native speaker has experience of very much more language than is contained in even the largest corpus, much of that experience remains hidden from introspection (although see Cook 2001). For example, native-speaker language teachers are often unable to say why a particular phrasing is to be preferred in a particular context to another, and the consequent rather lame rationale 'it just sounds better' is a source of irritation to learners.

Intuition is a poor guide to at least four aspects of language: collocation, frequency, prosody and phraseology. Examples of each of these are given below.

Judgements about collocations. Some collocations are easy to intuit (*play – game*, for example), others are more difficult. Granger (1998a) points out that some adverbs collocate with particular adjectives. She mentions 'stereotyped combinations such as *acutely*

aware, *keenly felt*, *painfully clear*, *readily available*, *vitally important*' which learners of English tend not to use (1998: 150), presumably because no course writer has had the accuracy of intuition to be aware of them. Johansson (1993: 46) adds several more common combinations, including *broadly comparable*, *comparatively new/small*, *deadly dull*, *deeply concerned*, *desperately worried*, *diametrically opposite*, *eminently respectable*, *equally good/important*, *exceptionally cheap/high*, *extremely difficult*, *fairly accurate/certain/small/wide*. It is difficult for the native speaker to be conscious of these combinations, and others like them, without corpus evidence.

Judgements about frequency. It is almost impossible to be conscious of the relative frequency of words, phrases and structures except in very general terms (anyone might guess that *take* is a more frequent verb than *disseminate*, but it is difficult to guess whether *fare* or *fantasy* is more frequent).[4] Halliday (1993: 3) points out that whereas people do have some intuitions about the frequency of lexis (*go* is more frequent than *walk*, which is more frequent than *stroll*), they are unlikely to have intuitions about the frequency of grammatical categories, and may even resist information about these.

Semantic prosody and pragmatic meaning. Channell (2000) makes the point very strongly that many instances of pragmatic meaning are beyond the reach of intuition. For example, she notes that the phrase *par for the course* is used not only to comment that something frequently happens, but also to evaluate that event negatively. Native speakers of English often react with surprise to information of this kind and sometimes attempt to find alternative explanations for it, such as a bias in the corpus. Often, though, the reaction is of surprised recognition – 'of course that's true, why didn't I think of it before?'.

Details of phraseology. Although native speakers can often recognise if a phraseology is unusual, articulating the nature of the atypicality may be more difficult. One example may be taken from Owen (1996). Owen notes that the Bank of English corpus has several examples of *require/s* followed by a passive to-infinitive clause, even though such a construction, as in *Further experiments require to be done*, seems wrong to Owen's native-speaker intuition. A closer look at phraseology resolves the problem. Although *REQUIRE to be* is found in the Bank of English, and fairly frequently, the past participle that follows is usually that of a verb with a specific meaning, not a general verb such as *do*. There are plenty of examples of the type *These roses require to be pruned each*

[4] In fact, *fare* occurs about 4,000 times in the Bank of English; *fantasy* about 10,000 times.

spring but very few indeed of *require to be done* (only 3 out of 302). Thus Owen's intuitions are backed up by the evidence of the corpus, but on phraseological rather than grammatical grounds. (See chapter 7 for more details of this argument.)

Although an over-reliance on intuition can be criticised, it would be incorrect to argue that intuition is not important. Indeed, it is an essential tool for extrapolating important generalisations from a mass of specific information in a corpus. An example can be taken from Sripicharn (1998). Looking in the Bank of English for examples of the verb and noun *CONTACT*, Sripicharn noticed that not only was the noun followed by *with* whereas the verb was not (the expected finding), but in addition that the verb was typically used with 'official' persons – an office, a newspaper etc (e.g. *Contact your local travel agent*) – whereas the noun was chosen when the person was a family member or friend (e.g. *She had no contact with her father*). Sripicharn intuitively realised that the difference between the two kinds of nouns (*travel agent* and *father*) was important.

Having argued for the benefits of corpora to the study of language, it is as well to consider also the limitations of a corpus. These might be summarised as follows:

1 A corpus will not give information about whether something is possible or not, only whether it is frequent or not. Descriptions of English are moving towards a concentration on the typical and away from notions of well-formedness (Sinclair 1991: 17; Biber et al 1998: 3), but questions of the type – 'Can I say this?' – still need to be answered. To give a specific example, is the verb *EXPIRE* used with the preposition *of*, by analogy with *DIE of* (e.g. *died of heart disease*)? The Bank of English shows us that *EXPIRE of* is rare: 5 lines out of 3,519 lines for *EXPIRE* (0.1%), compared with *DIE of*, which has 5,259 lines out of 85,511 lines for *DIE* (6%). It is also rare in comparison with *EXPIRE from* (12 lines). Furthermore, of the 5 lines of *EXPIRE of*, three are jokey, non-serious.[5] All this is useful information, but it does not actually answer the question as to whether *EXPIRE of* is acceptable English. Native-speaker intuition has to answer that question.
2 A corpus can show nothing more than its own contents. Although it may (justifiably) claim to be representative, all attempts to draw

[5] This example of a non-serious use is from a review of a film about poverty and revolution: *You know it will all end in tears, of course, with a lorry load of cast members either expiring of starvation, consumption or falling under the sharp heel of capitalism spurned.*

generalisations from a corpus are in fact extrapolations. A statement about evidence in a corpus is a statement about that corpus, not about the language or register of which the corpus is a sample. Thus conclusions about language drawn from a corpus have to be treated as deductions, not as facts.

3 A corpus can offer evidence but cannot give information. For example, what does the phrase *something of a* mean before a noun, as in *something of a surprise*? It might be supposed to be a 'downtoner': *something of a surprise* is a small surprise. Comparing *COME as something of a surprise* with *COME as a surprise* in the Bank of English affirms this supposition to some extent, in that adjectives such as *total*, *complete* and *big* occur before *surprise* but not in the phrase *something of a surprise*. Apart from that, however, there are no real clues as to the meaning of *something of a surprise* in the concordance lines. The corpus simply offers the researcher plenty of examples; only intuition can interpret them.

4 Perhaps most seriously a corpus presents language out of its context. The work of Kress and van Leeuwen (Kress and van Leeuwen 1994; Kress 1994), for example, depends upon a text being encountered in its visual and social context (or, more properly, the text consists not of the words alone but of the spatial context in which the words appear). A corpus (as corpora are currently conceived) cannot show this. Equally significant is the issue of spoken data, in that transcription can never represent intonation, kinesics ('body language'), and other paralinguistic information entirely accurately. Even if the issue of visual and intonational features is ignored, it remains true that a corpus masks some of the features of the texts in it by presenting concordance lines, in which the structure of the original is lost. These factors all show the need for a corpus to be one tool among many in the study of language.

Conclusion

This chapter has introduced some of the basic concepts that are important when using corpora in applied linguistics. Some of the possible uses of corpora have been demonstrated, along with arguments for the usefulness of corpora in describing how a language works and what language can show about the context in which it is used. Some of the limitations of corpus use have also been mentioned. The next three chapters will discuss in more detail the issues and methods in compiling and investigating corpora.

Note on sources of examples

Unless otherwise indicated, the examples, concordance lines and statistical information in this book come from the Bank of English corpus. Some of the concordance lines include codes used in that corpus, although many have been edited to exclude unnecessary codes. Examples of codes include:

	start paragraph
	end paragraph
<h>	start headline
</h>	end headline
<FO1>	female speaker number 1
<MO2>	male speaker number 2
<MOX>	unidentified male speaker

The concordancing program used ('Lookup') sometimes leaves incomplete words at the beginning or end of a concordance line.