1 Introduction

1.1 Introduction: the basics

Here we look at the basics of corpus linguistics, from what a corpus is to how to build one. We outline the basic functions of corpus software, such as generating word frequency lists and concordance lines of words and clusters (or chunks). We also try to give an idea of the wide range of applications of a corpus to fields as diverse as forensic linguistics and language teaching. Creating a corpus also brings up a number of issues, for example, whose language it is representing. This is particularly the case in relation to corpora of English in the context of native versus non-native speaker users of the language.

1.2 What is a corpus and how can we use it?

A corpus is a collection of texts, written or spoken, which is stored on a computer. In the past the term was more associated with a body of work, for example all of the writings of one author. However, since the advent of computers large amounts of texts can be stored and analysed using analytical software. Another feature of a corpus, as Biber, Conrad and Reppen (1998) point out, is that it is a *principled* collection of texts available for *qualitative* and *quantitative* analysis. This definition is useful because it captures a number of important issues:

A corpus is a principled collection of texts

Any old collection of texts does not make a corpus. It must represent something and its merits will often be judged on how representative it is. For example, if we decided to build a corpus representing classroom discourse in the context of English Language Teaching (ELT), how do we design it so as to best represent this? Would four hours of recordings from an intermediate level class in a London language school suffice? Great care is usually taken at the design stage of a corpus so as to ensure that it is representative. If we wished to build a corpus to represent classroom discourse in ELT, we would have to create a design matrix that would ideally capture all the essential variables of age, gender, location, type of school (e.g. state or private sector), level, teacher (e.g. gender, qualifications, years of experience, whether native or non-native speaker), class size (large groups, small groups or one-to-one), location, nationalities and so on. It is important to scrutinise how a corpus is designed when considering buying or accessing one, or when evaluating any findings based on it. The design criteria of a corpus allow us to assess its representativeness. Crowdy (1993), Biber (1993), McEnery and

Wilson (1996), McCarthy (1998), Biber, Conrad and Reppen (1998), Kennedy (1998), Meyer (2002), Thompson (2005a), Wynne (2005a), Adolphs (2006) and McEnery, Xiao and Tono (2006), among others, are essential reading if you are considering designing your own corpus.

A corpus is a collection of electronic texts usually stored on a computer

Because corpora are stored on a computer, this allows for very large amounts of text to be amassed and analysed using specially designed software. Language corpora can be composed of written or spoken texts, or a mix of both, and nowadays the capability exists to add multimedia elements, such as video clips, to corpora of spoken language. If it is a corpus of written language, texts may be entered into a computer by scanning, typing, downloading from the internet or by using files that already exist in electronic form.[1] For example, you may wish to build a corpus of your students' written work over a one-year period so as to track their vocabulary acquisition and to compare this with other data. This could be done easily by asking your students to email you their work (see section 1.4 for further details on creating your own corpus).[2] Corpora of spoken language, on the other hand, are much more time-consuming to assemble. For instance, if you wished to build a corpus of your own classroom interactions, you would first need to record the classes and then transcribe them. One hour of recorded speech usually yields approximately between 12,000 and 15,000 words of data and it takes around two days to transcribe, depending on the level of coding you decide to use in transcription (O'Keeffe and Farr 2003 discuss the pros and cons of building versus buying a corpus). For example, a spoken corpus may be coded for different speaker turns, interruptions, speaker overlaps, truncated utterances, extra-linguistic information such as 'giggling', 'door closes in background', 'dog barking' (see section 1.4). More detailed transcriptions include prosodic information as found in the London-Lund Corpus (Svartvik and Quirk 1980), the Lancaster/IBM Spoken English Corpus (Knowles 1990; Leech 2000) and the Hong Kong Corpus of Spoken English (Cheng and Warren 1999, 2000, 2002). Not surprisingly, written corpora are much more plentiful and usually much larger than spoken ones.

A corpus is available for qualitative and quantitative analysis

We can look at a language feature in a corpus in different ways. For example, using a corpus of newspapers, we could examine how many times the words *fire* and *blaze* occur. This will give us quantitative results, that is, numbers of occurrences, which we can then compare with frequencies in other corpora, such as casual conversation or general written English. This might lead us to conclude that the word *blaze* is more frequently used in newspaper articles than in general English conversation or writing, when talking about destructive outbreaks of fire. This conclusion is arrived at through quantitative means. However, another approach is to look more qualitatively at how a word or phrase is used across a corpus. To do this, we need to look beyond the frequency of the word's occurrence.

[1] It is essential to remember that most texts are covered by copyright, and that permission to use a text may need to be obtained before it can be stored or exploited in any way.

[2] Teachers may find that their institutions have strict ethical guidelines for using students' work in research, and these should always be observed.

As we will exemplify below, looking at concordance lines can help us do this and to see qualitative patterns of use beyond frequency.

1.3 Which corpus, what for and what size?

There is no one corpus to suit all purposes. The one we choose to work with is the one that best suits our needs at any given time. Begin with the question: *why do I need to use a corpus?* The answer to this question will vary widely. For example, some may wish to use a corpus for research purposes to study how a lexical item or pattern is used. Others may wish to compare the use of an item in different language varieties, for example *will* and *shall* in American versus British English (see Carter and McCarthy 2006: 880–881). In such cases, the corpus which is chosen must best represent the language or language variety, and, if comparing varieties, the corpora themselves must be comparable. For example, comparing *will* and *shall* in American and British English using a corpus of American academic textbooks from the 1960s and a corpus of contemporary spoken British English will obviously yield flawed results (unless one is conducting a study of language change and the possible backwash effects of spoken language on written language). In a pedagogic context, a corpus may also be utilised for reference purposes, for example, a teacher may advise students to search a corpus to find out what preposition most commonly follows *bargain* as a verb. Many of these types of questions can also be answered by looking things up in a dictionary. The advantage of looking up a lexico-grammatical query in a corpus is that it provides us with many examples of the search item in its context of use. However, a corpus will not tell us the meaning of the word or phrase. This is something that we have to deduce from the many examples that are generated. Combining a dictionary and a corpus can be a valuable route in a pedagogical context. Let us look the word *bargain* using a dictionary and some corpus examples:

Figure 1: Main entries for *bargain* from the *Cambridge Advanced Learner's Dictionary* (CD-ROM 2003)

bargain (AGREEMENT) /ˈbɑː.gɪn/ (US) /ˈbɑːr-/ *noun* [C]
an agreement between two people or groups in which each promises to do something in exchange for something else:
"I'll tidy the kitchen if you clean the car." "OK, it's a bargain."
The management and employees eventually ***struck/made*** *a bargain* (= reached an agreement).

bargain /ˈbɑː.gɪn/ (US) /ˈbɑːr-/ *verb* [I or T] (Verb Endings)
Unions bargain ***with*** *employers for better rates of pay each year.*
I realized that by trying to gain security I had bargained ***away*** *my freedom* (= exchanged it for something of less value).

bargain for/on *sth* *phrasal verb*
to expect or be prepared for something:
We hadn't bargained on such a long wait.
The strength of the opposition was rather ***more than*** *she'd bargained* ***for****.*

Figure 2: Sample of concordance lines for *bargain* from the Cambridge International Corpus (see Appendix 1 for details)

1	blic-sector unions have been allowed to	**bargain** away jobs for pay. In a deal
2	over ... The chancellor also asks us to	**bargain** away whatever obligations or int
3	: your loss is Southampton's gain. A	**bargain** buy at pounds 1 million this sea
4	weapons; and that the Russians will not	**bargain** for cuts in something that Labou
5	in his shirt front. Scurra has struck a	**bargain**,' he called out as he bustled fu
6	e, and even the possibility of making a	**bargain**,he turned his back on them for
7	tologists had kept to their side of the	**bargain**;he'd make their deaths quick...
8	he airport.' I see now why this is a	**bargain** holiday. Once the clients have p
9	erm these really s5 sort of quite	**bargain** holidays where you take+
10	Chuffed. You little	**bargain** hunter you. laughs
11	Events' are manna from heaven for the	**bargain** hunter. When shares get marke
12	ost of the phone calls I took were from	**bargain** hunters,' Steve says. While L
13	junkies, pop history freaks and casual	**bargain** hunters. Record Collector magazi
14	as keen on trail running as they are on	**bargain** hunting. A spokeswoman for PR co
15	and you'll lose a lot of wine into the	**bargain**. Reading a champagne label
16	point and got a little success into the	**bargain**, she'll go back to what she was
17	And it's invariably dishonest into the	**bargain**." So how has he managed to we
18	tanding but seem pretty boring into the	**bargain**. THERE was a moment about a t
19	t free tickets. He's a widower into the	**bargain**, they say. Quite a catch for som
20	ess accepted separate electorates and a	**bargain** was struck over the distribution
21	chaser and it really is if you like the	**bargain** we will strike and I like to thi
22	ents that they can actually strike up a	**bargain** with a patient. Em and things ca
23	occurred to me that I might be able to	**bargain** with him. If you really are a Ke
24	es." But you're not. All you have to	**bargain** with now is a copy of the decode
25	added. The Americans are prepared to	**bargain** with the Russians on almost anyt
26	ers from their beds each day at five to	**bargain** with the wholesalers, which g

As well as illustrating a range of prepositions that follow *bargain*, the concordance lines also give a rich insight into how the word collocates with other words (see below and chapter 2), for example, *to strike a bargain*, or *bargain hunters*. We also find idiomatic usage, such as *into the bargain* meaning 'as well'.

On the question of corpus size, in the case of *bargain*, we had to search over 10 million words of data to find a range of instances. This is because it is not a core vocabulary item in English. If, on the other hand, we were looking at a word or structure that was quite common, a smaller corpus would suffice. Aston (1997), Maia (1997) and Tribble (1997) suggest using a small corpus if we are dealing with a very specialised language register, for words of caution, see Gavioli (2002) (see also chapter 8 which makes a case for using small corpora to look at relational language). In terms of what constitutes a large or a small corpus, it depends on whether it is a spoken or written corpus and what it is seeking to represent. For corpora of spoken language, anything over a million words is considered to be large; for written corpora, anything below five million is quite small. In terms of suitability, however, it is often the design of a corpus as opposed to its size which is the determining factor. For example, a corpus containing only highly technical engineering language will be largely inappropriate for language teacher trainees wanting to investigate general vocabulary. Therefore, while size is an issue, it should be considered hand-in-hand with the appropriateness of corpus design (for further discussion of these and other issues relating to size and corpus design see: Sinclair 1991a; Thomas and Short 1996; Aston 1997; Maia 1997; Tribble 1997; Biber et al. 1998; McCarthy 1998; Biber et al. 1999; Coxhead 2000; Carter and McCarthy 2001; Hunston 2002; O'Keeffe and Farr 2003; Thompson 2005a; Wynne 2005a; Adolphs 2006 and McEnery et al. 2006).

Overview of existing corpora

There are many corpora available and some can be bought, some are free and some are not publicly available (e.g. corpora compiled by publishers for the specific commercial purposes of producing language teaching resources and materials, or corpora where the consent agreement of writers or speakers may only allow for restricted use). Appendix 1 provides an overview of a wide range of language corpora and how to find out more about them. Throughout this book we will be referring to a number of these corpora in our illustrations and analyses.

1.4 How to make a basic corpus

A basic language corpus can be assembled from spoken or written texts and can be used with commercially available corpus software such as *Wordsmith Tools* (Scott 1999) and *Monoconc Pro* (2000), which any average home computer user can manipulate with relative ease. A spoken corpus takes considerably longer to build, as discussed above, because speech has to be transcribed and possibly coded for some of its non-verbal features. Written corpora, on the other hand, can be made very quickly using the internet as a source (though international copyright must always be respected in the usual ways).

Stages of building a spoken corpus

1 Create a design rationale

Your corpus will need some design principle (see above on representativeness). When considering the design of a spoken (or written corpus), considerations of feasibility (what is available, what is ethical, what is legal?) will need to be a guiding factor also. Decide what it is you wish to represent and consider how best you can represent this for your purposes. This will guide your decision as to how much data you want to collect. For example, you might wish to create a corpus of news reports to use in class. You could decide to collect ten news reports or a hundred. You may wish to only record business reports or political reports and so on.

2 Record data

It is useful to keep in mind that one hour of continuous everyday, informal conversation yields approximately 12,000 to 15,000 words. The mode of recording is also worth consideration. There are a number of options including analogue cassettes, digital media and audiovisual digital recorders. Traditional analogue, though they are inexpensive, have a number of drawbacks. They are cumbersome to store and unlike digital recordings, they cannot easily be computerised and aligned with the transcription later. Using digital devices leaves open the option of aligning sound (and image if you use an audiovisual recorder) with your transcription. Permission to record should be cleared in advance with the speakers and consent forms should be signed off authorising the use of the recordings for research or commercial pedagogical materials, etc. It may be necessary to specify how

the recordings will be used when obtaining permission; for example, is the speaker signing permission just for the transcript to be used, or for his/her actual voice to be used in research or any publication?

3 Transcribe recordings and save as text files

Spoken data needs to be manually transcribed and this is what makes corpora of spoken language such a challenge. They are best stored as 'plain text' files, as this offers the maximum flexibility of use with different software suites. As mentioned above, every one hour of recorded speech can take approximately two working days to transcribe. In most cases, every word, vocalisation, truncation, hesitation, overlap, and so on, is transcribed, as opposed to a cleaned-up version of what the speakers said. The level of detail of the transcription is relative to the purpose of your corpus. If you have no requirement to know where overlapping utterances and interruptions occur, then there is no point in spending time transcribing to that level of detail. Figure 3 shows an example of an extract from a transcript from the Limerick Corpus of Irish English (LCIE) (see appendix 1). Our data extracts in this book will use these conventions to a greater or lesser extent:

TRANSCRIPTION CODING KEY

<$1>, <$2>, etc.	these mark the different speakers in the order in which they appear on the recording
+	interruptions can be marked from where they occur and from where the utterance is resumed (often called 'latched turns')
=	unfinished or truncated words can be marked, for example, yester=
<?>	unintelligible utterance
<$E> laugh <\$E>	extralinguistic information such as 'laughing', 'sound of someone leaving the room', 'coughing', 'dog barking' can be useful background information

Figure 3: Extract of a transcript of a recording of family members changing a printer cartridge while looking at the instruction manual (from LCIE)

```
<$1> Oki Jet. Isn't that what we have?

<$2> Yeah but that's not the <$E> pause one second <\$E> there's a <?>. Here it is.
     Here Brendan. Here. Look. <$E> intercom goes off in the kitchen <\$E>

<$1> Knock that off now. <$E> sound of intercom being switched off <\$E>

<$2> There's about six different languages.

<$1> So what's the problem?

<$2> We needed to replace the print head.

<$1> Oh right.

<$2> So that's the problem. <$E> noise of printer in background <\$E>

<$3> <$E> shouting from another room <\$E> Hello.

<$2> <$E> looking at printer manual <\$E> Changing the ink cartridge <?>

<$3> <$E> from the other room <\$E> Change the+

<$1> Changing the ink cartridge yeah. What does it say abou=

<$2> Open the printer cover.

<$1> All right.

<$2> <$E> reading from the instruction manual <\$E> The print head carriage will move
      automatically to the head loading replacement position of the empty print head.

<$1> Right.

<$2> <$E> reading from the instruction manual <\$E> Release only the ink cartridge
      from the print head casing pulling gently outwards the lateral+

<$1> Press the green button first Brian

<$2> That's the black one. No that's fine. If you put that back in+

<$1> There's no print head on it.
```

4 Database texts

Transcription files need to be organised so that source information can be traced. For example, it may be useful to be able to retrieve information such as gender, age, number of speakers, place of birth, occupation, level of education, where the recording took place, relationship of speakers and so on. This information can be stored at the beginning of each transcript as an information 'header' (see Reppen and Simpson 2002: 98–99), or in a separate database, where the information is logged with the file name.

5 Check transcription

Finally, the transcription needs to be checked with the original recording for accuracy.

Stages of building a written corpus

1 *Create a design rationale*

As discussed above, start with a design rationale. Decide what it is you want to represent and how many texts you need to do this, from how many sources and over what period.

2 *Input texts*

Depending on what form they are in, written texts may need to be re-typed or scanned. They may already be in electronic format or may be downloadable from the internet, and may have special copyright restrictions on their use. Once they are in electronic form, they need ideally to be saved as 'plain text' files; once again, this will offer the maximum flexibility of use with different software suites.

3 *Database texts*

Any individual text in a corpus needs to be traceable to its source information (that is, who wrote it, where and when it was published, genre, number of words and so on, especially for purposes of subsequent use in relation to copyright). As discussed above, this can be stored at the beginning of each file (as 'header information') or in a separate database.

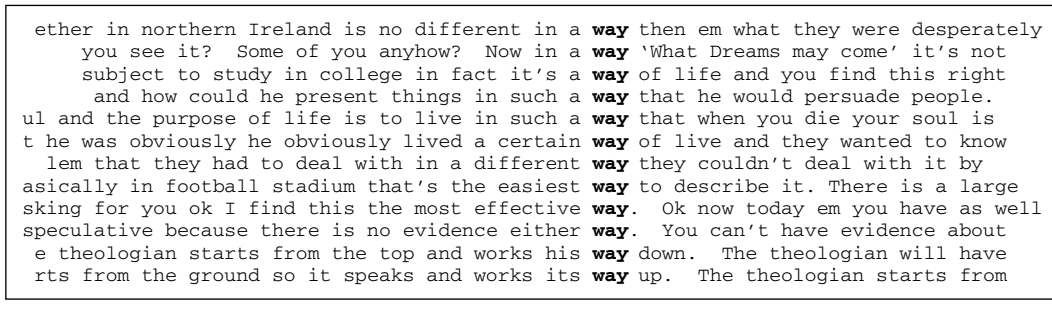
1.5 Basic corpus linguistic techniques

Here we overview some of the basic techniques that can be used on a corpus, using standard software such as *Wordsmith Tools* (Scott 1999) and *Monoconc Pro* (2000). Applications of these techniques will be illustrated throughout the book.

Concordancing

Concordancing is a core tool in corpus linguistics and it simply means using corpus software to find every occurrence of a particular word or phrase. This idea is not a new one and many scholars over the years have manually concordanced the Christian Bible, for example, painstakingly finding and recording every example of certain words. With a computer, we can now search millions of words in seconds. The search word or phrase is often referred to as the 'node' and concordance lines are usually presented with the node word/phrase in the centre of the line with seven or eight words presented at either side. These are known as Key-Word-In-Context displays (or KWIC concordances). Concordance lines are usually scanned vertically at first glance, that is, looked at up or down the central pattern, along the line of the node word or phrase. Initially, this may be disconcerting because we are accustomed, in Western cultures, to reading from left to right. Concordance lines challenge us to read in an entirely new way, vertically, or even from the centre outwards in both directions. Here are some sample lines from a concordance of the word *way* using the Limerick Corpus of Irish English (LCIE):

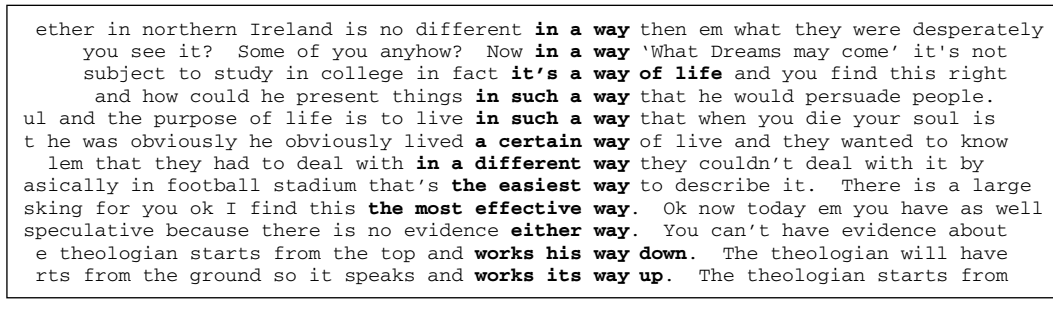
Figure 4: Concordance lines for *way* from LCIE



```
 ether in northern Ireland is no different in a way then em what they were desperately
     you see it?  Some of you anyhow?  Now in a way 'What Dreams may come' it's not
     subject to study in college in fact it's a way of life and you find this right
      and how could he present things in such a way that he would persuade people.
ul and the purpose of life is to live in such a way that when you die your soul is
t he was obviously he obviously lived a certain way of live and they wanted to know
  lem that they had to deal with in a different way they couldn't deal with it by
asically in football stadium that's the easiest way to describe it. There is a large
sking for you ok I find this the most effective way.  Ok now today em you have as well
speculative because there is no evidence either way.  You can't have evidence about
 e theologian starts from the top and works his way down.  The theologian will have
 rts from the ground so it speaks and works its way up.  The theologian starts from
```

Most software allows the number of words at either side of the node word or phrase to be adjusted to allow more of the context to be viewed and you can usually go back very easily and quickly to the source file containing the full text or transcript. Software normally facilitates the sorting of the concordance lines so that we can examine the lexico-grammatical patterns which occur before and/or after the node word. When sample concordance lines for *way* are sorted alphabetically to the left of the screen for example the following patterns emerge:

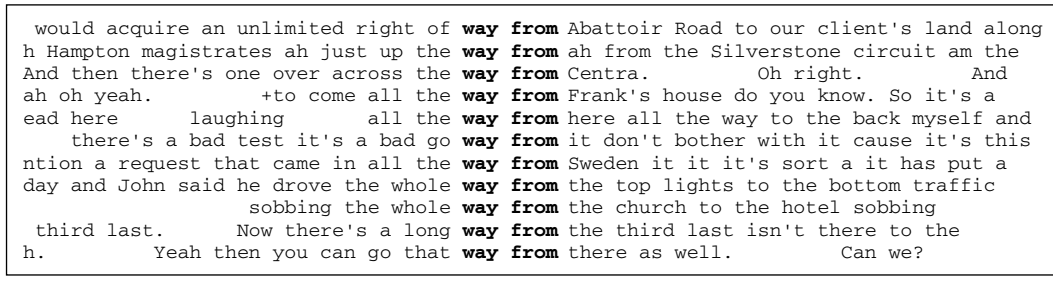
Figure 5: Sample concordance lines for *way* from LCIE, sorted to the left of the screen



```
 ether in northern Ireland is no different in a way then em what they were desperately
     you see it?  Some of you anyhow?  Now in a way 'What Dreams may come' it's not
     subject to study in college in fact it's a way of life and you find this right
      and how could he present things in such a way that he would persuade people.
ul and the purpose of life is to live in such a way that when you die your soul is
t he was obviously he obviously lived a certain way of live and they wanted to know
  lem that they had to deal with in a different way they couldn't deal with it by
asically in football stadium that's the easiest way to describe it.  There is a large
sking for you ok I find this the most effective way.  Ok now today em you have as well
speculative because there is no evidence either way.  You can't have evidence about
 e theologian starts from the top and works his way down.  The theologian will have
 rts from the ground so it speaks and works its way up.  The theologian starts from
```

Another random sample from the concordance lines of the word *way*, sorted to the right of the screen, shows a systematic pattern with *from*:

Figure 6: Sample concordance lines for *way* from LCIE, sorted to the right of the screen



```
 would acquire an unlimited right of way from Abattoir Road to our client's land along
h Hampton magistrates ah just up the way from ah from the Silverstone circuit am the
And then there's one over across the way from Centra.          Oh right.          And
ah oh yeah.          +to come all the way from Frank's house do you know. So it's a
ead here      laughing       all the way from here all the way to the back myself and
    there's a bad test it's a bad go way from it don't bother with it cause it's this
ntion a request that came in all the way from Sweden it it it's sort a it has put a
day and John said he drove the whole way from the top lights to the bottom traffic
                  sobbing the whole way from the church to the hotel sobbing
 third last.      Now there's a long way from the third last isn't there to the
h.         Yeah then you can go that way from there as well.          Can we?
```

Because concordance lines can provide many examples of patterns of use, they have application to the language classroom and are now being used in ELT materials. For example, here is an extract from the entry on *there* in *Natural Grammar* (Thornbury 2004: 155), where concordance lines have been adapted for an inductive grammar task:

Figure 7: Extract from *Natural Grammar* (Thornbury 2004: 155)

Exercises

1 Look at these concordance lines, and identify the meaning of *there* in each case. Is it a pronoun (showing that something exists) or is it an adverb (saying where something is)?

a **There**'s a bar and a lecture room for guests' use.
b **There**'d been another quake at 4am, a 6.5 shock.
c It was only in my third year that I really felt happy **there.**
d You say **there**'s a certain amount of risk. How much?
e I was **there** for her birth and it was the most exciting thing.
f But **there**'ll be no alcohol on sale.
g He was standing **there** with Mrs Kasmin as she tried to give him tea.
h He had been **there** since he left the Pit a year earlier.
i He was confident **there**'d be no problem. So was I.

Another example is found in McCarthy and O'Dell (2002), where students are invited to look at an extract from a concordance for the word *eye* and to decide which of the occurrences are idiomatic/metaphorical.

Figure 8: Extract from *English Idioms in Use* (McCarthy and O'Dell 2002: 109)

50.4 Here are some random examples from a computer database containing lines from real conversations. The figures in diamond brackets, e.g. <s1>, <s2>, mean 'first speaker', 'second speaker', etc. How many of the examples use *eye* as an idiom, and how many use the word *eye* in its literal sense as 'the organ we see with'? Use a dictionary if necessary.

1	go into town and get erm an	**eye**	test. <s1> Mm. <s2 > In town.
2	you er keep an	**eye**	out for tramps, do you then?
3	In your mind's	**eye**	how are you going to do that?
4	<s1> So I'll keep a general	**eye**	on it. And er <s3> Yeah
5	<s1> There's something in my	**eye**.	There's that thing floating
6	difficult to put that to your	**eye**.	You also have to have one eye
7	good offer? <s2> Yeah it caught my	**eye**	<s1> Yeah it's
8	I'm casting my	**eye**	over this form and I think
9	this year. <s4> Just keep an	**eye**	out for it. <s4> Yeah.
10	<s2> You'll have to keep an	**eye**	on her. <s1> Yeah. <s2> Oh my
11	so you're about	**eye**	level with the monitor.
12	saw her out of the corner of my	**eye**.	<s3> Her lipstick is all over

Word frequency counts or wordlists

Another common corpus technique which software can perform is the extremely rapid calculation of word frequency lists (or wordlists) for any batch of texts. By running a word frequency list on your corpus, you can get a rank ordering of all the words in it in order of frequency. This function facilitates enquiry across different corpora, different language varieties and different contexts of use. Below, for example are the first ten words from five different corpora (see appendix 1):

Table 1: Comparison of word frequencies for the ten most frequent words across five different datasets

	1	2	3	4	5
Rank order	**Shop (LCIE)**	**Friends (LCIE)**	**Academic LIBEL**	**Australian Corpus of English**	**CIC newspaper & magazine sub-corpus**
	spoken	**spoken**	**spoken**	**written**	**written**
1	you	I	the	the	the
2	of	and	and	of	to
3	is	the	of	and	of
4	thanks	to	you	to	a
5	it	was	to	a	and
6	I	you	a	in	in
7	please	it	that	is	is
8	the	like	in	for	for
9	yeah	that	it	that	it
10	now	he	is	was	that

1 Service encounters: a sub-corpus of the Limerick Corpus of Irish English (LCIE) consisting of shop encounters (8,500 words)
2 Friends chatting: a sub-corpus of LCIE, consisting of female friends chatting (40,000 words)
3 Academic English: The Limerick-Belfast Corpus of Academic Spoken English (LIBEL CASE, one million words of academic English[3])
4 Australian casual conversation: the Macquarie Corpus of English (ACE) (one million words of written Australian English)
5 Written British and American English: The Cambridge International Corpus based on a 100,000 word sample of newspaper and magazines from McCarthy (1998: 122–123).

[3] Hereafter, LIBEL CASE will be referred to as LIBEL.

Even from just the first ten words of these corpora, tendencies emerge in terms of genres and contexts of use. The shop (column 1) and casual conversation (column 2) results show markers of interactivity typical of spoken English such as *I, you, yeah* (as a response token), *like, please* and *thanks* (see Carter and McCarthy 2006). Though the academic corpus (column 3) is also naturally-occurring speech, the first ten words lack the interactive markers found in the first two columns. The academic corpus results resemble more the written data from the ACE and CIC (columns 4 and 5). All three share features associated with written language, that is to say the high frequency of:

- articles *a* and *the*, indicating a high instance of noun phrases
- the preposition *of*, suggesting post-modified noun phrases
- *that*, especially in academic corpora, pointing to its multi-functionality, as a subordinator (particularly following report verbs or in *it* patterns) as well as as a relative pronoun in relative clauses
- prepositions *to, for* and *in*, suggesting prepositional phrases

Conversely, there is a lack of:

- interactive pronouns *I* and *you*; the only pronoun that figures in the top ten words is *it*, which is referential as opposed to interactive
- response tokens or discourse markers such as *yeah, like, now*

In a number of chapters in this book we will use word frequency lists. In chapter 2 for example, word frequencies will form the basis for identifying the core vocabulary of English for pedagogical purposes in identifying different target levels.

Key word analysis

This function allows us to identify the key words in one or more texts. Key words, as detailed by Scott (1999), are those whose frequency is unusually high in comparison with some norm. Key words are not usually the most frequent words in a text (or collection of texts), rather they are the more 'unusually frequent' (ibid). Software compares two pre-existing word lists and one of these is assumed to be a large word list which will act as a reference file or benchmark corpus. The other is the word list based on the text(s) which you want to study. The larger corpus will provide background data for reference comparison. For example, we saw above that *the* is the most frequent word in the LIBEL corpus of spoken academic English (table 1); if we select one economics lecture from this corpus and generate a word list, we can also see that *the* is again the most frequent word. However, if we compare this economics lecture word list with the larger one from the LIBEL corpus using keyword software (such as that found in *Wordsmith Tools*), it will tell us which words occur with unusual frequency, or 'keyness'. These words are then referred to as the key words.

Table 2: Key words from an economics lecture relative to a general corpus of academic lectures

1	tax	15	higher
2	income	16	percent
3	system(s)	17	rates
4	average	18	ordinary
5	basic	19	sixty
6	rate	20	marginal
7	supply	21	scheme
8	poor	22	labour
9	thousand	23	terms
10	impact	24	cost(s)
11	equity	25	characterised
12	under	26	workers
13	both	27	systems
14	figures	28	negative

Scott (1999) notes the key word facility provides a useful way of characterising a text or a genre and has potential applications in the areas of forensic linguistics, stylistics, content analysis and text retrieval. In the context of language teaching, it can be used by teachers and materials writers to create word lists, for example in Languages for Specific Purposes programmes (e.g. English for pilots, French for engineers), where the key specialised vocabulary can be automatically identified, either from a single text (e.g. an aeronautical training manual) or from a corpus of specialised texts.

Cluster analysis

As chapters 2 and 3 will illustrate, the analysis of how language systematically clusters into combinations of words or 'chunks' (e.g. *I mean, this that and the other*, etc.) can give insights into how we describe the vocabulary of a language. It also has implications for what we teach in our vocabulary lessons and how learners approach the task of acquiring vocabulary and developing fluency. As a corpus technique the process of generating chunks or cluster lists is similar to making single word lists. Instead of asking the computer to rank all of the single words in the corpus in order of frequency, we can ask it to look for word combinations, for example 2-, 3-, 4-, 5-, or 6-word combinations (for further explanation of how this works, see chapter 3). By way of example, using *Wordsmith Tools*, table 3 shows the 20 most frequent 3-word combinations from 10 million words (five million written and five million spoken) of the Cambridge International Corpus (CIC):

Table 3: The 20 most frequent three-word chunks in 10 million words from CIC

	Chunk	Frequency per million words		Chunk	Frequency per million words
1	I don't know	588	11	a couple of	166
2	a lot of	364	12	do you want	159
3	one of the	320	13	you have to	158
4	I don't think	248	14	be able to	157
5	it was a	240	15	a bit of	155
6	I mean I	220	16	you want to	153
7	the end of	198	17	and it was	148
8	there was a	193	18	it would be	142
9	out of the	190	19	do you know	138
10	do you think	177	20	you know what	137

Chapter 3 looks in detail at chunks in spoken and written corpora and at the pedagogical implications of these patterns.

1.6 Lexico-grammatical profiles

A further corpus strategy, when looking at concordance lines, is to create a 'lexico-grammatical profile' of a word and its contexts of use. A lexico-grammatical profile describes typical contexts in terms of:

1 Collocates: which word(s) occur most frequently and with statistical significance (i.e. not just by random occurrence) in the word's environment?
2 Chunks/idioms: does the word form part of any recurrent chunks? Is the word idiom-prone? What types occur (for example, binomials or trinomials such as *rough and ready*, or *ready, willing and able*)?
3 Syntactic restrictions: are there syntactic patterns which restrict the word? For example, are there prepositions that go with the word? What are its typical clause-positions (initial/medial/final)? Are there any tense/aspect restrictions?
4 Semantic restrictions: are there semantic restrictions? For example, the word/phrase is applied to humans only, or is never used with an intensifier.
5 Prosody: 'Semantic prosody' is a term used by Louw (1993) and means simply that words, as well as having typical collocates (for example, *blonde* typically collocates with *hair*, but not with *car*), tend to occur in particular environments, in a way that their meaning, especially their connotative and attitudinal meanings, seem to spread over several words. For example, words might tend to occur overwhelmingly in positive or in negative environments. Stubbs (1995), for instance, shows how more than 90% of the collocates of *cause* are negative, for example *accident,*

cancer, commotion, crisis and *delay*. By way of a positive semantic prosody example, he offers *provide*, which typically collocates with, for example, *care, food, help, jobs, relief* and *support*. Before the advent of computerised language analysis, this phenomenon had never been properly codified in terms of actual usage.
Another example of prosody is seen in the CIC data for the adjective *prim*, where the word seems strongly associated with old-fashioned, frumpy, conservative, mostly female attributes. Figure 9 shows a sample concordance for *prim*.

6 Other relevant or recurring features.

Figure 9: Concordance for *prim* (CIC, 10 million words mixed spoken/written)

```
1       stuff of sensible office suits and prim 50s ensembles, dogtooth is
2              You're too       You're too prim and proper to sit in the
3  girls.          No. But this one's real prim and proper and oh you know
4      . The young today are not nearly so prim and proper as we were.
5      o me.     Mm.         So English so prim and proper in the way he
6      stands either. Mum taught us. We're prim and proper. Mandy, fired up
7      ed his father-in-law's picture of a prim galleon on frilly sea.
8      re."    Hallo," said Alma, thin and prim, in a hurry. They must be
9       delightful part of my life in-that prim, incongruous little parl
10     , thinks you should leave alone the prim little fork that's always
11     day that she died she was a star. A prim Miss Marple lookalike in
12       she blushed furiously feeling all prim. '' So it's powerful stuff t
13      ness,' Bless replied, now sounding prim. The stranger smiled at him
14     roof. Anna thought it looked like a prim woman with its neat apron
15     at their tender leaves. This small, prim woman, devoted to Professor
```

A lexico-grammatical profile is principally drawn from concordance lines, though the frequency and keyness of any item in a particular corpus may also be of relevance. The lines should ideally be sorted and analysed in both screen directions, left and right. Figure 10 (overleaf) shows an example for the word *abroad* using the framework we have just outlined. A lexico-grammatical profile for *abroad*, based on figure 10, would give us the following: Left-screen sorting seems to produce the most visible and productive patterning since *abroad* tends to be phrase-, clause- or sentence-final.

1 Collocates of three or more occurrences: *be, been, go, trip, travel, work*.
2 Chunks/idioms: *home and abroad* occurs three times.
3 Syntax: *abroad* only seems to be used adverbially; no preposition after verbs of motion (*flow, go, shift, travel*); no preposition after *trip/holiday*; only one preposition occurs (*from*). It can be used as a post-nominal modifier (*trip abroad, holidays abroad*).
4 Semantics: *abroad* can be used with static or dynamic verbs; it is never premodified (for example, **very abroad*, **far abroad* do not occur). Its most frequent meaning is geographical or political, but there are also examples where it simply means 'in the public domain/out in the open' (lines 30, 33, 39).
5 Prosody: *abroad* is anywhere, not the writer's country or the country in question, often in contrast to the UK or 'home', a place to which people travel for leisure and work and where trade and investment are seen as important; no particular connotations of negativity, but sometimes a prosody of 'difference' or 'exoticness' (lines 15, 22, 25, 48, 49).

Figure 10: Random sample of 60 concordance lines for *abroad*, based on five million words of mixed written texts (CIC)

1	iaspora continues with their activities	**abroad**	In the relatively low-tech car i
2	to ease the curbs on travel at home and	**abroad**.	If the reforms really take hold
3	ervices, to firm leadership at home and	**abroad**;	to conditions in which business
4	s examples of good practice at home and	**abroad**.	Local Authority Involvement
5	which attracts adults from Ireland and	**abroad**	to courses in Donegal on Irish l
6	well-managed forests in both the UK and	**abroad**.	FSC-certified charcoal is so
7	deal route for visitors from the UK and	**abroad**.	It is easily accessible by rail
8	n West Germany, 60 % of whose sales are	**abroad**,	has no foreigner on board. Elec
9	companies (about 70 % of its sales are	**abroad**),	makes all its Walkmans in Japa
10	new of the murder, although he had been	**abroad**	when it had taken place. The
11	s for the day. The younger son being	**abroad**,	I sent him the news with a litt
12	aking place. The flows of Japanese cash	**abroad**,	mainly across the Pacific, are
13	our responsibility to our own citizens	**abroad**,	is not an easy question. We can
14	direct investment by Canadian companies	**abroad**.	An example of the first was the
15	ay. He also wore a bored, Englishman-	**abroad**	look that suggested he might rat
16	y can get around the rules is to expand	**abroad**	rather than at home. Industriali
17	n spent at home to raise incomes flowed	**abroad**	instead. Japan's government,
18	weatshops. Even designs are coming from	**abroad**	- from 'cheap' fashion centres l
19	vertising standards, are delivered from	**abroad**	every week. The bill is adding t
20	l is being sent into British homes from	**abroad**	- and it is subsidised by the Po
21	g in enormous amounts of hot money from	**abroad**	by offering high interest to pay
22	s first overseas trip. I would never go	**abroad**,	because I'd always heard the ba
23	at deal about it. It means he has to go	**abroad**	a lot. He's in Paris at the mome
24	espectfully and indigenously. If you go	**abroad**	this summer, support the local c
25	they're fed up with the hassle of going	**abroad**,'	said Stan, executive member of
26	hey didn't suffer because she was going	**abroad**.	It all took her longer than
27	t of the chamber of commerce, have gone	**abroad**	to avoid arrest. General Noriega
28	y mum and dad. It was our first holiday	**abroad**	and we went to Majorca. There wa
29	want" Andrew explains. Regular holidays	**abroad**	are also affordable. Florida is
30	on appeared to be the only living human	**abroad**	at that ungodly hour. When sa
31	here. About 14 % of JVC's production is	**abroad**,	up from 9 % in 1985. JVC's fina
32	ed how many of last year's 183 journeys	**abroad**	were necessary. They included
33	y, £20 ## There's a big lie	**abroad**	and it's about taxes and the wel
34	retire at 30. She has no plans to live	**abroad**,	as Morceli has done (in Califor
35	to Amy Johnson, but both are now living	**abroad**	and, although they have been con
36	sts. Success also means selling more	**abroad**.	A Russia no longer losing groun
37	to be the case then I'd probably move	**abroad**.	But that would only happen if I
38	e caused by the book caused her to move	**abroad**,	first to New Mexico where she e
39	. A new, deficit-induced realism is now	**abroad**.	This week a draft report by the
40	nds. Who commands the purse, at home or	**abroad**?	That cohabitation did not me
41	itish embassies and other organizations	**abroad**,	gathering intelligence in place
42	ed for minimum cover (i.e. third party)	**abroad**.	ADVENTURE and high risk spor
43	the inmates choose to write to penpals	**abroad**.	Tito has been writing to a penp
44	eek before he was due to take up a post	**abroad**	as a correspondent for a western
45	jewellery boxes which he tries to sell	**abroad**.	He also spends a lot of time "t
46	pub with a soldier while I was serving	**abroad**	I'd give her such a pasting she
47	technologies will eventually be shifted	**abroad**	- but not until the factories no
48	Disorientated and thinking he was still	**abroad**,	he shouted: `I'm English like y
49	s checking up on the way they do things	**abroad**,'	explained his wife Mavis. T
50	port is to make it easier when I travel	**abroad**.	Apart from that, I consider
51	ertainly mean that he will never travel	**abroad**	again, and inevitably both he an
52	national decline until he had travelled	**abroad**	and discovered that, far from be
53	ier in the month I'd made my first trip	**abroad**	and came up against another set
54	ay for too long now. His frequent trips	**abroad**	had become a fact of her life bu
55	he Children, in the course of her trips	**abroad**;	these are located around the bu
56	after six months she resigned and went	**abroad**.	Years of exile followed, in Mal
57	Kitty, with Jefferson and Edwina, went	**abroad**	for a few months to escape atten
58	apply for a licence for minors to work	**abroad**.	That continued until I was eigh
59	who leave the country intending to work	**abroad**	for more than a year are deemed
60	there had been mention of a son working	**abroad**,	but it had been a long time ago,

1.7 How have corpora been used?

Lexicography

Language corpora have many applications beyond language description for its own sake. They are now the standard tool for lexicographers, who use multi-million word corpora to examine word frequency, patterning and semantics in the compilation of dictionaries. This tradition of basing dictionary entries on actual use rather than intuition is not entirely new. In the 1700s, when Samuel Johnson was compiling the first comprehensive dictionary of the English language, he manually collated a corpus of language based on samples of usage from the period 1560 to 1660. Three centuries later, the corpora that lexicographers use are vast, methodical collections of both spoken and written texts; at the time of writing, the Cambridge International Corpus (CIC) has over one billion words. They are constantly added to and facilitate the monitoring of language trends and usage changes. Some publishers also hold learner corpora, for example the CIC consists of over 27 million words of learner writing, 12 million of which are error coded. This provides very useful information about the types of lexical and grammatical errors that are made and in so doing allows for dictionary writers and other materials writers to highlight typical problems. The pioneering work in this area was the *Collins Birmingham University International Language Database* (COBUILD) project. This was set up at the University of Birmingham in 1980 under the direction of John Sinclair. To date it has produced 16 dictionaries and grammars, most influentially the *Collins COBUILD English Language Dictionary* (1987, 2nd edition 1995, 3rd edition 2001, 4th edition 2003) and the *Collins COBUILD Grammar Patterns* series (1996; 1998). It also sparked the design of the Lexical Syllabus (see Willis 1990). All major publishers now provide corpus-based dictionaries.

Grammar

The COBUILD project also had a major influence on grammar. It provided the concept of 'pattern' as an interface between lexis and grammar. How 'pattern grammar' emerged through corpus-based lexico-grammatical research, the debates which surrounded it and its application for language teaching are covered extensively in Hunston and Francis (2000), see also Hunston et al. (1997). Major grammars of English are now corpus-informed (for example, Quirk et al. 1985; Sinclair 1990; Biber et al. 1999; Carter and McCarthy 2006). In recent years, Biber et al. (1999) conducted a seven-year grammar project which led to the creation of their corpus-based grammar of English. It focuses on American and British English and on the four registers of conversation, fiction writing, news writing, and academic writing. This grammar was based on the analysis of a 40 million word corpus of spoken and written texts. Carter and McCarthy (2006) based their grammar on the CIC, at that time consisting of over 700 million words of English, constructed over a ten-year period and still in the process of development. It includes examples from sources such as newspapers, best-selling novels, non-fiction books on a wide range of topics, websites, magazines, junk mail, TV and radio programmes and recordings of people's everyday conversations in a variety of social settings ranging from university seminars

to intimate family conversations. Carter and McCarthy found that it was crucially important in many cases to separate statements made about spoken as opposed to written grammar, and include a CD-ROM where users can access sound-clips for the more than 7,000 example sentences and utterances recorded in the grammar, in the belief that spoken grammar especially needs to be heard and not just read from a page. As in the case of lexicography, corpora have revolutionised how grammar is studied. Corpus tools allow grammarians to extensively investigate grammatical frequency and patterning, to look in detail at differences in the use of grammar in different varieties of language, and readily provide contemporary examples of actual language usage. By attesting structures and patterns across a wide range of speakers and social and geographical contexts (using the database information referred to above for features such as age, gender, educational background, etc.), Carter and McCarthy were able to include features in widespread spoken usage, even though they may be frowned upon by traditionalists (see also Carter 1999b, 2005). In chapters 5 and 6, we look at how corpus-based grammar has forced us to distinguish between patterns which can be viewed prescriptively (for example that third-person singular present-tense verbs end in -s) and patterns that are less fixed and need to be viewed probabilistically (we provide a detailed case study of the *get*-passive structure to exemplify this in chapter 5).

Stylistics

In other language-related fields, corpora are also being used. In the area of stylistics, for example, which is mostly concerned with the study of the language of literature, Burrows (2002) notes that traditional and computational forms of stylistics have much in common. Both rely upon the close analysis of texts, and both benefit from opportunities for comparison. According to Wynne (2005b) corpus linguistics is opening up new vistas for stylistics, and there are interesting similarities in the approaches of stylistics and corpus linguistics. Stylistics, he notes, is a field of empirical inquiry, in which the insights and techniques of linguistic theory are used to analyse literary texts, that is by applying systems of categorisation and linguistic analysis to, for example, poems and prose (see van Peer 1989; Leech and Short 1981; Louw 1993; Short 1996; Short et al. 1996; Semino et al. 1997; Semino and Short 2004). A related area of increasing interest in the study of language and literature is the notion of 'semantic prosody' (Louw 1993), which we mentioned earlier in relation to lexico-grammatical profiling. Wynne (2005b) tells us several corpus linguists have used evidence of these patterns to study creativity in language, both in fiction and in everyday usage (Sinclair 1987a, 1987b; Carter 2004; Hoey 2005; Stubbs 2005). The work of Louw is of particular importance for the study of stylistics. His important 1993 paper comes from the lineage of J. R. Firth and Sinclair; it provides a novel methodology for analysing literary texts through the study of collocations, based on the idea that certain words, phrases and constructions become associated with certain types of meaning due to their regular co-occurrence with the words of a particular semantic category (for a more recent survey see Wynne 2005b).

Translation

Language corpora have considerable application in the area of translation (see Teubert 1996, 2002; Tognini-Bonelli 1996; Zanettin 1998, 2002; Claridge 2000; Serpollet 2002). As noted by Aston (1999), this has been from two main perspectives, descriptive and practical; that is to say descriptive research which looks at corpora of translations, comparing these with corpora of original texts so as to establish the characteristics both peculiar and universal to translated texts (Gellerstam 1996; Baker 1995, 1998; Laviosa 1998). On the other hand, Aston observes, corpora have been looked at as aids in the processes of human and machine translation, and for this purpose he distinguishes between three main types of corpora:

Monolingual corpora

These consist of texts in a single language, which may be either the source or the target language of a given translation.

Comparable corpora

Where monolingual corpora of similar design are available for two or more languages, they may be treated as components of a single comparable corpus. Baker (1995) suggests that comparable corpora have the potential to reveal most about features specific to translated text.

Parallel corpora

These also have components in two or more languages, consisting of original texts and their translations, for example, a novel and its translation in another language. Aston (1999) points to the distinction between 'unidirectional parallel corpora' which consist of texts in one language along with translations of those texts into another language (or languages) and 'bidirectional' or 'reciprocal parallel corpora' which contain four components: source texts in language A and their aligned translations in language B, and source texts in language B and their aligned translations in language A. Parallel corpora exist for several language pairings including English–French (for example, Church and Gale 1991; Salkie 1995), English–Italian (Marinai et al. 1992), and English–Norwegian (Johansson and Hofland 1994; Johansson et al. 1996). Typical applications of parallel corpora include translator training, bilingual lexicography and machine translation.

For further reading about the use of translation corpora see, for example, Johansson and Hofland (1994); Johansson and Ebeling (1996); Sinclair et al. (1996); King (1997); Laviosa (1998); Santos (1998); Salkie and Oates (1999); Santos and Oksefjell (1999); Altenberg and Granger (2002); Salkie (2002); Van Vaerenbergh (2002), among others.

Forensic linguistics

Another area which is increasingly using language corpora as a tool is forensic linguistics, which broadly concerns itself with the use of language in law and crime investigation. Corpora have many applications relative to the diversity of the focus of the discipline itself, which includes the analysis of the genuineness of documents from confessions to suicide notes, authorship identification in academic settings (e.g. issues of plagiarism), ransom

notes, threat letters, readability/comprehensibility of legal language, forensic phonetics (e.g. speaker identification), police interview and interrogation data, language rights of ethnic minorities, as well as the discourse of the courtroom setting (see for example Gibbons 1994, 2003; Conley and O'Barr 1998; Shuy 1998; Tiersma 1999; Cotterill 2002a, 2002b, 2003, 2004; Heffer 2005; Tiersma and Solan 2005). Corpora can be used to look at large amounts of courtroom data; for example, Cotterill (2002b) used a corpus of the entire internationally notorious O. J. Simpson trial in the United States. Corpora can be used to compare language patterns; for example, Boucher (2005), in his analysis of features of deceit in recounting, compared a corpus of 200 three- to five-minute discourses where half represented truthful and half inaccurate accounts. He was able to statistically describe significant differences in variables such as hesitation, lexical repetition and utterance length. Authorship and plagiarism are growing concerns within forensic linguistics, for which corpora can prove a useful instrument of investigation (see Coulthard 2004; Solan and Tiersma 2004).

Sociolinguistics

Corpora have also had an impact in the area of sociolinguistics. Their application in this area is not surprising given that many corpora of spoken language, in particular, can be built around sociolinguistic variables such as age, gender, level of education, socio-economic background and so on. Regional variation, for example, can be explored using language corpora. Ihalainen (1991a) looked at variation in verb patterns in south-western British English, while Ihalainen (1991b) compared the grammatical subject in educated and dialectal English in the London-Lund and the Helsinki Corpus of modern English dialects. Kirk (1992, 1999) and Kallen and Kirk (2001) look at languages in contact in the context of Northern Ireland and Irish English, Ulster Scots, Irish and Scots Gaelic using a corpus-based approach. The SCOTS corpus (see Douglas 2003, Corbett and Douglas 2004) offers great potential for sociolinguistic study. It aims to represent the present-day linguistic situation in Scotland eventually representing written and spoken data of Scottish English and Scots, Scots Gaelic as well as non-indigenous community languages such as Punjabi, Urdu and Chinese (see appendix 1).

Age-related research is prevalent especially in the context of teenager language. The Bergen Corpus of London Teenage Language (COLT) (see Haslerud and Stenström 1995; Stenström 1998; and Appendix 1) has provided the basis for numerous studies. Features such as discourse markers have been given particular attention; for example, Andersen (1997a, 1997b) focuses on the use of *like* in London teenage speech. The use of tags is linked to age in a number of studies (Stenström 1997a; Stenström et al. 2002). Hasund (1998) looks at class-determined variation in the verbal disputes of London teenage girls, while Hasund and Stenström (1997) examine conflict talk using a corpus-based comparison of the verbal disputes of adolescent females. Other corpus-based studies on language and gender include Aijmer (1995) which looks at apologies, Holmes (2001) which examines linguistic sexism and Mondorf (2002), a study of gender differences in English syntax.

Taboo language is also looked at using corpora such as COLT and the British National Corpus (see Stenström 1995; Stenström et al. 2002; and Appendix 1). Corpus-based sociolinguistic studies that look at non-standard usage include Stenström (1997b), which

again focuses on London teenager usage. Callahan (2004) explores Spanish-English code switching using a corpus comprised of 30 fictional works from 24 Latino authors published in the United States, between 1970 and 2000. Callahan shows that written codeswitching follows for the most part the same syntactic patterns as its spoken counterpart. Her corpus findings also point to the use of non-standard English, which appears in 53% of the corpus in the forms of African-American Vernacular English and certain varieties of New York English. Lapidus and Otheguy (2005), in another New York corpus-based study, look at language contact in the context of English and Spanish. They focus on the use of non-specific *ellos* (English equivalent: *they*). One of Lapidus and Otheguy's main conclusions is that the susceptibility of language varieties to contact influence is primarily at the discourse-pragmatic level. Corpora have had a major influence in the areas of discourse and pragmatics also and throughout this book we will draw on examples of such work.

1.8 How have corpora influenced language teaching?

As we discussed above, the processes of dictionary-making have been revolutionised by the use of language corpora and this obviously feeds into language teaching materials. All major learners' dictionaries of English are now based on constantly updated multi-million word databases of language. Fundamentally, corpora have provided evidence for our intuitions about language and very often they have shown that these can be faulty when it comes to issues such as semantics and grammar. As we noted earlier, we now increasingly base our major grammars, like dictionaries, on large language corpora. The contribution of corpus linguistics, therefore, to the description of the language we teach is difficult to dispute. According to McCarthy (2001: 125) corpus linguistics represents cutting-edge change in terms of scientific techniques and methods and probably foreshadows even more profound technological shifts that will 'impinge upon our long-held notions of education, roles of teachers, the cultural context of the delivery of educational services and the mediation of theory and technique'.

As well as providing an empirical basis for checking our intuitions about language, corpora have also brought to light features about language which had eluded our intuition (e.g. the frequency of ready-assembled chunks; see chapter 3). In terms of what we actually teach, numerous studies have shown us that the language presented in textbooks is frequently still based on intuitions about how we use language, rather than actual evidence of use. While there are often sound pedagogical reasons for using scripted dialogues, their status as a vehicle for enhancing conversation skills has been challenged in recent years (Carter 1998; Burns 2001; Burns, Joyce and Gollin 2001; McCarthy and O'Keeffe 2004; Thornbury and Slade 2006). Burns (2001) notes that scripted dialogues rarely reflect the unpredictability and dynamism of conversation, or the features and structures of natural spoken discourse, and argues that students who encounter only scripted spoken language have less opportunity to extend their linguistic repertoires in ways that prepare them for unforeseeable interactions outside of the classroom. Holmes (1988: 40), for example, looked at epistemic modality in ESL textbooks as compared with corpus data and found that many textbooks devoted an

unjustifiably large amount of attention to modal verbs, at the expense of alternative linguistic strategies. Boxer and Pickering (1995) showed contrast between speech acts in textbook dialogues with real spontaneous encounters found in a corpus. Carter (1998) compares real data from the Cambridge and Nottingham Corpus of Discourse in English (CANCODE, see appendix 1) with dialogues from textbooks and finds that the dialogues lack core spoken language features such as discourse markers, vague language, ellipsis and hedges. Gilmore (2004) examines the discourse features of seven dialogues published in course books between 1981 and 1997, and contrasts them with comparable authentic interactions in a corpus. He finds that the textbook dialogues differ considerably from their naturally-occurring equivalents across a range of discourse features including turn length and patterns, lexical density, number of false starts and repetitions, pausing, frequency of terminal overlap or latching, and the use of hesitation devices and response tokens. He looks at dialogues from more recent course books and finds that there is evidence that they are beginning to incorporate more natural discourse features. The *Touchstone* series (McCarthy, McCarten and Sandiford 2005a and b, 2006a and b) is an attempt to show how course book dialogues, and even entire syllabi, can be informed by corpus data. In addition to the conventional four-skills syllabus strands of speaking, listening, reading and writing, the *Touchstone* authors provide a syllabus of conversational strategies, based on the most common words and phrases in the North American spoken segment of the CIC. The strategies recur throughout the four levels of the multi-skills programme and are graded. An example is given in figure 11, where the discourse marker *I mean* is exploited.

Figure 11: Extract from the *Touchstone* series (McCarthy, McCarten and Sandiford 2005a: 49)

A Complete the questions or answers with your own ideas. Compare with a partner. Do you have any of the same ideas?

1. *A* Do you ever go out after class?
   B Well, not very often. I mean, I usually go straight home .
2. *A* How do you like the restaurants in your neighborhood?
   B They're not bad. I mean, they're _____ .
3. *A* Are you busy in the evening? I mean, do you _____ ?
   B Well, I take a lot of classes.
4. *A* What do you do in your free time?
   B Well, I don't have a lot of free time. I mean, _____ .

About you → **B** ***Pair work*** Ask and answer the questions. Give your own answers.

Kettemann (1995) highlights the mismatch between actual language use and the prescription often found in pedagogical grammars that reported speech involves the 'backshift rule' for tenses in the reported speech constructions (see also Baynham 1991, 1996; McCarthy 1998). Hughes and McCarthy (1998) look at the use of past perfect verb forms and find that, across a wide range of speakers in the CANCODE corpus, the past perfect has a broader and more complex function in spoken discourse than hitherto described. Corpus descriptions have also enhanced our understandings of units of fixed phrasing, collocation, and more extended language patterns (Sinclair 1991a, 2003a, 2004; Svartvik 1991; Aston 1995; McCarthy and Carter 2002; Biber et al. 2004; Schmitt 2004; Thornbury and Slade 2006). Throughout the chapters that follow, we will survey and build on relevant findings from corpus research and tease out the implications these have for language teaching.

Corpora of learner languages are a relatively recent, but very important development. Granger (2003), a forerunner in the area, defines a learner corpus as an electronic collection of authentic texts produced by foreign or second language learners. She notes that, in the early 1990s, publishers and academics started, independently but concurrently, to gather and analyse learner data. The International Corpus of Learner English (ICLE, see Granger 1993, 1994, 1996, 1998a; Granger et al. 2002), initiated around that time, currently contains over two million words of writing by learners of English from 19 different mother tongue backgrounds. The writing in the corpus (essays) has been contributed by advanced learners of English as a foreign language rather than as a second language and is made up of 19 distinct sub-corpora, each containing one language variety (English to French, English to German, English to Swedish, etc.). This corpus is error-coded, which allows for invaluable research into typical learner error patterns (see Dagneaux et al. 1996; De Cock et al. 1998). Findings from research into learner corpora can be addressed in materials design, including the development of Computer Assisted Language Learning (CALL) applications. For example, Altenberg and Granger (2001), looking at Swedish- and French-speaking learners, examine the use of high frequency verbs, and in particular use of the verb *make*. As well as looking at the role of transfer in the misuse of these verbs relative to native-speaker norms, they investigate whether learners tend to over- or underuse these verbs and whether high frequency verbs are error-prone or safe. They find that EFL learners, even at an advanced proficiency level, have great difficulty with high frequency verbs such as *make*. They suggest that concordance-based exercises (see Data-driven learning below) can help raise awareness of the complexity of high frequency verbs. Learner spoken data have also been collected, a notable example being the Louvain International Database of Spoken English Interlanguage (LINDSEI) set up in 1995 (see De Cock 1998, 2000). This provides spoken data for the analysis of the speech of second language learners (see also Granger et al. 2002). Numerous other studies have been conducted using learner corpora, including Granger (1996, 1997, 1998a, 1998b, 1998c, 1999, 2002, 2003, 2004), De Cock and Granger (2004), Meunier (2002a, 2002b), Gilquin (2003) and Cosme (2004).

Data-driven learning

Computer Assisted Language Learning (CALL), among many other applications, includes the use of language corpora, where learners get hands-on experience of using a corpus through guided tasks or through materials based on corpus evidence, such as concordance lines on handouts (see Johns 1991a). Here an inductive approach relies on an 'ability to see patterning in the target language and to form generalisations' about language form and use (Johns 1991a: 2). This activity is commonly referred to as 'data-driven learning' (DDL) after Johns (1986 and 1991a). Johns (2002: 108) sees DDL as a process which 'confront(s) the learner as directly as possible with the data', 'to make the learner a linguistic researcher' where 'every student is Sherlock Holmes'. Over the years Johns, among others, has developed the idea and contributed many teaching materials based on the DDL approach (see Johns 1988, 2002; Stevens 1991; Wichmann 1995; Fox 1998; Kettemann 1995; Tribble and Jones 1990; 1997; Flowerdew 1993, 1996; Gavioli 1996; Wichmann et al. 1997; Tribble 2000, 2003; Aston 2001). A basic internet search will bring up numerous homepages dedicated to DDL, which provide many useful links to resources (such as online corpora and concordancers), research findings and materials. Such a search is also evidence of the popularity of DDL among language teachers, many of whom post their materials online and conduct action research into the classroom application of these materials. DDL, like corpus linguistics in general, is not without its critics (see Widdowson 1991, 2000; Prodromou 1996, 1997a, 1997b; Owen 1996; Seidlhofer 1999; Bernardi 2000; see below for further discussion of issues and debates). Many also question the application of DDL to lower-level learners, though some studies provide evidence of its use at lower levels (see Johns 1988, 2002; St John 2001; Kennedy and Miceli 2002).

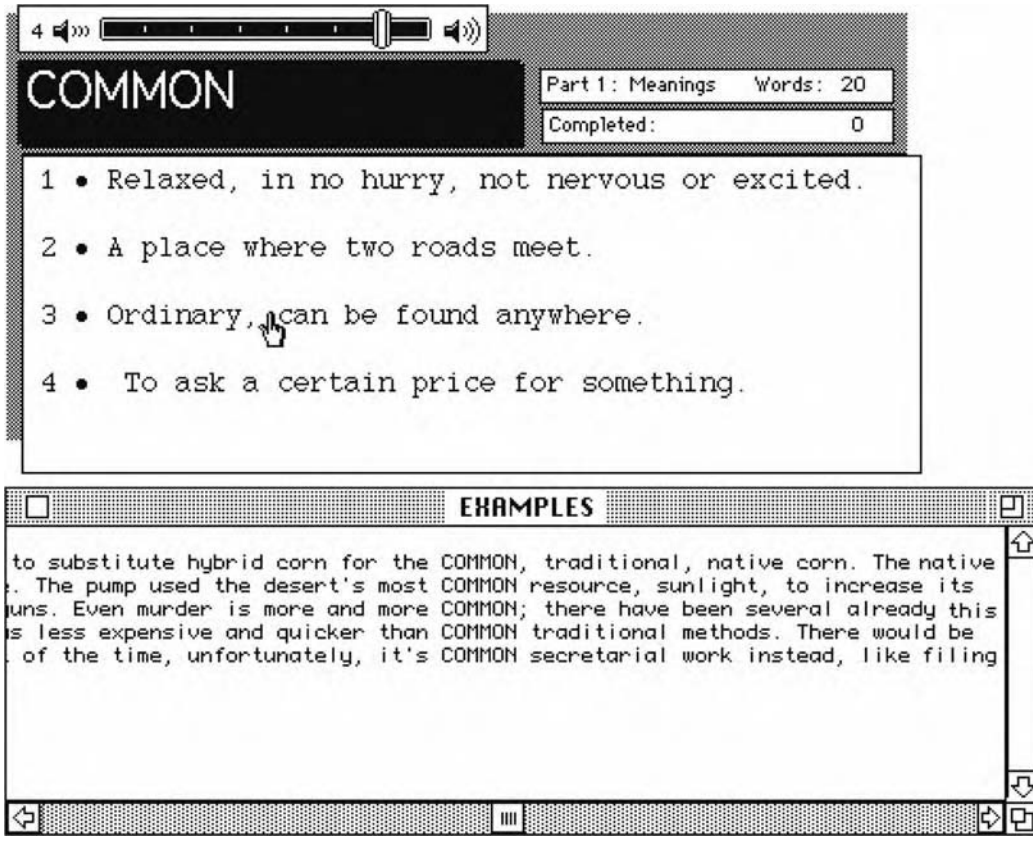
Chambers, who has been involved in the development of a one-million word corpus of journalistic French (see appendix 1: Chambers-Rostand Corpus of Journalistic French; Chambers and Rostand 2005), provides a number of illustrations of how DDL can be used in the context of teaching French and how it can facilitate the development of learner autonomy (see Chambers and Kelly 2002, 2004; Chambers and O'Sullivan 2004; Chambers 2005; Braun and Chambers 2006; Chambers in press; O'Sullivan and Chambers in press). Chambers and Kelly (2002) note that the pedagogical context of DDL brings together constructivist theories of learning, the communicative approach to language teaching and developments within the area of learner autonomy. Cobb (1997) points to the potential of DDL to provide multiple contextual encounters for the acquisition of new vocabulary. The literature on vocabulary acquisition, according to Cobb, is virtually unanimous on the value of learning words through several contextual encounters (Mezynski 1983; Stahl and Fairbanks 1986; Krashen 1989; Nation 1990). Language learners are advised to read more (see Krashen 1989) so as to facilitate multi-contextual lexical acquisition. In reality, Cobb notes that few language learners have time to do enough reading for natural, multi-contextual lexical acquisition. DDL may have a role in rationalizing and shortening this learning process by providing a rich source of embodiments and contexts from new vocabulary. Empirical studies on the learning benefits of DDL are relatively few, but they do show positive results (see for example Cobb 1997; Turnbull and Burston 1998; Kennedy and Miceli 2001; Lenko-Szymanska 2002). Cobb (1997) reports on his longitudinal study of vocabulary

acquisition using concordance line tasks. This study provides interesting examples (with screen shots) of a variety of sequential DDL activities which draw on a specially designed corpus of 10,000 words (comprised of 20 texts of about 500 words each, assembled from the students' reading materials). Figure 12 shows the opening task:

Figure 12: Example of DDL task from Cobb (1997)

Part 1: Choosing a meaning. The learner is presented with a small concordance of four to seven lines, in KWIC format with the to-be-learned word at the centre, and uses this information to select a suitable short definition for the word from one correct and three randomly generated choices.

1 Choosing a meaning



(Cobb 1997, available online http://www.er.uqam.ca/nobel/r21270/cv/Hands_on.html).

1.9 Issues and debates in the use of corpora in language teaching

Authenticity of materials for language teaching and learning

As we have seen, collecting data for use in a corpus means collecting examples of language as it is actually used in authentic contexts. Debate over the extent to which authentic

language should form the basis of language courses has been taking place for the last thirty years or so (Canale and Swain 1980; Breen 1983; Van Lier 1996; Rost 2002) but it has been re-energised by the availability of corpus data.

It is often argued that, in language teaching, examples drawn from corpus sources should form the basis for the material used to exemplify the language and that an aim of language teaching should be to produce learners who are able to communicate effectively and competently. In order for this to happen, it is argued further, learners need to experience authentic rather than contrived examples of data; by 'contrived' is meant examples of language that are specially made up or invented for the pedagogic purposes of illustrating a particular feature or rule of the language. One problem is that the terms 'contrived' and 'authentic' have become emotionally charged and in opposition to each other.

The availability of corpus examples has produced a different perspective since we can find in corpora numerous examples of texts that are free-standing, in so far as they are independent of any language learning task. They are in their own authentic context, and they are composed for a particular audience (which tends to be different to that of the language learner). Thus, when they are presented with corpus examples, learners encounter real language as it is actually used, and in this sense it is 'authentic'. However, the language has been wrenched from its original context, and so, in one sense, is 'decontextualised'. This position suggests that as soon as texts are extracted from the context in which they first appeared, are stored in large electronic databases, and are reproduced for the teaching context, they are effectively removed from an authentic environment. The learner, then, has to process such texts with reference to a different context than the one in which they originated, a context which may not reflect his or her communicative goals in the classroom context. Furthermore, one can argue that authentic texts are embedded in particular cultures and may thus be culturally opaque to those outside that (usually western) culture, and that it may, as a result, be next to impossible for learners to 'authenticate' such texts for themselves on this basis. Authenticity should therefore preferably be defined as a relationship between a text and the response that it triggers in its immediate audience (see for example Lee, 1995; Widdowson 1996, 1998). Consequently, there is among many a preference for contrivance and the deliberate use of culturally 'neutral' examples as a more solid basis for a pedagogy that is sensitive to learners' needs. Such contrived texts also allow for material to be more easily graded for learners at different levels of competence. Another non-corpus-based option is to use texts suggested or provided by the learners themselves, which will, by definition, be potentially maximally authentic.

Supporters of the view that there should be more authentic material available in classrooms argue, on the other hand, that naturally-occurring data can be carefully chosen and mediated, that it can be contextualised for the learner, that learners are no different from other human beings, who have a natural proclivity to contextualise language data for themselves, and that the use of such data in the classroom can actually facilitate discussion of cultural background, as well as provide more grounded motivation because the text is so obviously a 'real' example of the target language (Peacock 1997). To deprive learners of such experiences for ideological reasons without consulting them is,

in the opinion of the present authors, patronising and self-defeating. Others advance a related argument that tasks can be graded according to the nature of the authentic material (Willis and Willis 1996; Bygate et al. 2001; Willis 2003). The latter position would also seem to be an argument for a more careful pedagogic selection of materials from authentic sources. In our experience, corpora, both spoken and written, do indeed contain many texts that are obscure and culturally opaque, but they also contain numerous texts that are transparent, easily contextualised and interpretable by any mature human being. It is simply a matter of how carefully one selects the material, who the end-users are and what they want and expect from a language programme. For centuries, language teachers have plucked written texts out of the contexts in which they were originally produced and imported them into the classroom, carefully selecting and mediating them for their students; we see the use of corpora in this connection as an example of historical continuity which harnesses the technical possibilities of speeding up searches for useful and usable material. Many teachers are now using the world's biggest corpus, the internet, and its associated search engines, in just this way.

These issues are addressed in several places in this book. Our basic position is that for most pedagogic purposes in most contexts of teaching and learning a language, it is preferable to have naturally-occurring, corpus-based examples than contrived or unreal examples, but always in the context of freedom of choice and careful mediation by teachers and/or materials writers who know their own local contexts. For further reading on the debate that surrounds this see Sinclair (1991a, 1991b), Aston (1995), Carter and McCarthy (1995), Prodromou (1996), Owen (1996), Carter (1998), Cook (1998), Seidlhofer (1999), Widdowson (2000, 2001).

The 'native speaker' and the classroom

Authentic language invariably invokes the idea of language drawn from sources supplied by native speakers and recent research has shown that language learners often regard the approximation to native speaker English as a main goal in the language learning process (Timmis 2002). While the notion of the native speaker of English tends to be used to refer to those whose first language is English, the concept is a complex one (Roberts 2005), as there are, as Rampton (1990) and others have demonstrated, non-native speakers who have great affiliation to a language and are more competent in that language than native speakers. The vast number of different varieties of 'native speaker' English (e.g. American, British, Irish, Australian, South African, Singaporean) means that this notion cannot easily be translated, or modelled, into one particular standard for the language classroom, although international publishers tend to focus on either American or British English as a model.

Whether we are referring to contrived, invented or naturally-occurring samples of English, the choice of a particular variety for the ELT context, even down to fine-grained choices of a particular regional or local variety, is inevitably to some degree a matter of ideology and invariably a political issue. At the same time, it is acknowledged that the proportion of English exchanged daily between non-native speakers is growing rapidly, with an overall increase in globalisation and internationalisation (see Crystal 1997) to the point

where non-native users of English far outnumber native speakers of English (Graddol 1998), undermining, for some, any privileging of native speaker discourse.

At the same time this raises the further question whether native-speaker models are the most appropriate basis for language learners, who may predominantly use their L2 to operate in an international, rather than a 'native' context. This state of affairs has led some to propose that English as a Lingua Franca (ELF) is more significant internationally than English as a first or second language and that consequently, corpora of non-native Englishes are needed in order to help us identify the kinds of English crucial to communication in such ELF contexts (see below) and to use such evidence as a preferred basis for classroom teaching and learning (see Medgyes, 1994; Braine 1999; Oda 1999, 2000; Jenkins 2000; Tajino and Tajino 2000; Seidlhofer 2001a; Carter and Fung (forthcoming) for further discussion on native versus non-native speaking teachers).

ELF: English as a lingua franca

Seidlhofer (2001a: 143–4) notes that while learner corpora (see above) have their use as a 'sophisticated tool for analysing learner language . . . some of the data in the learner corpora could also contribute to a better understanding of English as a lingua franca'. Seidlhofer goes on to detail a corpus development which she has championed: The Vienna-Oxford International Corpus of English (VOICE), a collection English as a Lingua Franca (ELF) currently under construction. Here lingua franca is defined as an additionally acquired language system that serves as a means of communication for speakers from different speech communities, who use it to communicate with each other but for whom it is not their native language. It is 'a language which has no native speakers' (Seidlhofer 2001a: 146) (see also Malmkjær 1991; House 1999, 2002, 2003; James 2000). The initial target for the VOICE corpus is to collect around half a million words of spoken data from speakers whose first language is not English and whose primary and secondary education did not take place in English, but who make use of English as a lingua franca (ELF) (see Seidlhofer 2004). In a parallel development, Mauranen (2003) reports on a corpus of ELF in academic settings (EFLA) at the Tampere Technology University, Finland. Its initial target is to collect half a million words of spoken data from two university settings. Both Seidlhofer and Mauranen aim, through empirical investigations of ELF, to show that a sophisticated and versatile form of language can develop which is *not* a native language (Seidlhofer 2001b; Mauranen 2003). Seidlhofer (2001a) argues that this is a much-needed development to fill the conceptual gap between the growing recognition and meta-linguistic discussions about global English and the existence of a codified form which eventually might have pedagogical applications in the identification of the most efficient forms of communication in the domain of ELF. With this in mind, the corpus may establish 'something like an index of communicative redundancy' (Seidlhofer 2001a: 147). Early findings from the VOICE corpus (see Seidlhofer 2004) tentatively identify a number of features which point to systematic lexico-grammatical differences between native-speaker English and ELF, for example dropping the third person present tense 's' (e.g. *she look*), omitting definite and indefinite articles, insertion of prepositions (e.g. *can we discuss about this issue*). These features often

involve typical errors which most English teachers would correct and remediate. However, Seidlhofer points out that they appear to be generally unproblematic and do not cause an obstacle to communicative success in ELF. The work of Jenkins (1996, 2000, 2004, 2005) has also been very influential here in relation to the teaching of pronunciation for ELF. She makes a parallel argument relating to ELF phonology. Her research finds that a number of items common to most native-speaker varieties of English were not necessary in successful ELF interactions; for example, the absence of weak forms in words like *from* and *for*; and the substitution of voiceless and voiced *th* with /t/ or /s/ and /d/ or /z/ (e.g. *think* became *sink* or *tink*, and *this* became *dis* or *zis*). Jenkins argues that such features occur regularly in ELF interactions and do not cause intelligibility problems.

Developments in and findings from corpus-based ELF studies further the debate about 'ownership' and function of a language like English and their empirical findings put forward ELF as a pedagogical model which challenges the accepted native-speaker-based norms of EFL. However, great uncertainties remain in this area, not least whether the object of description is a *function* of English rather than a codifiable variety, that is to say a way in which people adapt differently to every different circumstance and make greater or lesser use of their communicative repertoire depending on the exigencies of each individual interaction. Mauranen (2003) confidently labels ELF as a variety, but much discussion is still needed as to what, exactly is meant by 'variety' here. Other problems arise in the (perhaps unfair) equation between a reduced or 'stripped down' ELF syllabus and an impoverished experience of the L2. Indeed, it could be argued that learners of any language always end up producing less than the input they are exposed to, and that if that input itself is deliberately restricted, then even less will be the outcome, and so on. Lastly, the evidence so far as to what exactly ELF is is rather scant, and there is reason to believe that East Asian ELF, for example (e.g. a Chinese speaker interacting in English with a Korean speaker) may be very different from European ELF (e.g. a Danish speaker using English with a Dutch speaker) and we may need to describe many 'ELFs' to get anywhere near an accurate picture of the global uses of English. What the present authors do support, however, is the way native-speaker corpora of spoken language, with all their attendant shortcomings, have sparked a lively if sometimes heated debate as to the most suitable models of English for pedagogy. This is a step forward from the days when southern-England, middle-class English was unquestioned as the pedagogical model in most parts of the world (the situation which pertained when two of the present authors began their teaching careers). We also support the move to build more and yet more useful corpora from a wider range of different settings.

SUEs or Successful Users of English

Rather than continuing to focus solely on the native speaker, we should begin to look much more closely at the notion of the 'expert user' and at ideas advanced by Prodromou and others (Prodromou 2003a, 2005) concerning what he terms SUEs (or Successful Users of English). As we discuss in chapter 4, Prodromou (2005) takes idiomaticity as a paradoxical example of something which, for native speakers, makes life easy, enabling fluent production

of deeply culturally-embedded chunks heard and rehearsed since childhood. These same idiomatic chunks seem to place impossible obstacles in the path of the non-native speaker, however proficient. SUEs are highly successful L2 communicators, but they will achieve this goal by strategic use of their resources in ways different from those of native speakers. It makes more sense, therefore, not to see SUEs as failed native speakers, but to look upon all successful users of a language, whether native- or non-native-speaking, as 'expert users'.

A spoken corpus can underline for us how important it is to look closely at what speakers and listeners do, whoever those speakers are, whether they are native or non-native. Such research shows that our ability to interact with others is an important part of what makes us successful users of the language and is, we believe (and this is confirmed by research that is reported throughout this book), what learners of English aspire to know about and do in and with a language, and for the very reason that they know that this is what they do successfully in their first language. We will never meet those needs just by introspecting on what we *think* we say, nor by feeding our learners an impoverished diet of what we think they need based on those intuitions; only by respecting learners' and teachers' choices and aspirations within their own local contexts will we best serve them.

When we do look at what speakers and listeners do, we may not hear native speakers as we might want to hear them or as how we might have learned to expect to hear them. But we do hear real people interacting with one another, working at full stretch with the language, adjusting millisecond by millisecond to the interactive context they are in, playing with the language, being creative, being affective, being interpersonal and, above all, expressing themselves as they engage with the processes of communication which are most central to our lives. It is hard to imagine any learner of a second language not wanting to be a good, human communicator in that second language, whether they are going to use it with native speakers or with any other human beings. Language teaching can only benefit from even closer inspection of such fundamentally human processes. And the road from corpus to pedagogy, upon which we take tentative, sometimes faltering steps in this book, is an essential part of that process.