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