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...

**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)

<table>
<thead>
<tr>
<th>bargain (AGREEMENT)</th>
<th>/ˈbɑːɡ(ə)n/</th>
<th>noun [C]</th>
</tr>
</thead>
</table>
| 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).

<table>
<thead>
<tr>
<th>bargain</th>
<th>/ˈbɑːɡ(ə)n/</th>
<th>verb [1 or T] (Verb Endings)</th>
</tr>
</thead>
</table>
| 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).

<table>
<thead>
<tr>
<th>bargain for/on sth</th>
<th>phrasal verb</th>
</tr>
</thead>
</table>
| 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.
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; Adolfs 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 spo-
ken 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 tran-
scription 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 <$\textbackslash$E>$ extralinguistic information such as ‘laughing’, ‘sound of
someone leaving the room’, ‘coughing’, ‘dog barking’ can be
useful background information
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):
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 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 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 t he was obviously he obviously lived a certain *way* that when you die your soul is 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 asking for you ok I find this the most effective speculative because there is no evidence either 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 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 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 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 asking for you ok I find this the most effective speculative because there is no evidence either 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 and then there's one over across the ah oh yeah. +to come all the read here. Laughing all the 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 day and John said he drove the whole sobbing the whole *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 guest’s use.
   b. There’d been another quake at 4am, a 6.8 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.

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