1 What is corpus linguistics?

1.1 Introduction

What is corpus linguistics? It is certainly quite distinct from most other topics you might study in linguistics, as it is not directly about the study of any particular aspect of language. Rather, it is an area which focuses upon a set of procedures, or methods, for studying language (although, as we will see, at least one major school of corpus linguists does not agree with the characterisation of corpus linguistics as a methodology). The procedures themselves are still developing, and remain an unclearly delineated set – though some of them, such as concordancing, are well established and are viewed as central to the approach. Given these procedures, we can take a corpus-based approach to many areas of linguistics. Yet precisely because of this, as this book will show, corpus linguistics has the potential to reorient our entire approach to the study of language. It may refine and redefine a range of theories of language. It may also enable us to use theories of language which were at best difficult to explore prior to the development of corpora of suitable size and machines of sufficient power to exploit them. Importantly, the development of corpus linguistics has also spawned, or at least facilitated the exploration of, new theories of language – theories which draw their inspiration from attested language use and the findings drawn from it. In this book, these impacts of corpus linguistics will be introduced, explored and evaluated.

Before exploring the impact of corpora on linguistics in general, however, let us return to the observation that corpus linguistics focuses upon a group of methods for studying language. This is an important observation, but needs to be qualified. Corpus linguistics is not a monolithic, consensually agreed set of methods and procedures for the exploration of language. While some generalisations can be made that characterise much of what is called ‘corpus linguistics’, it is very important to realise that corpus linguistics is a heterogeneous field. Differences exist within corpus linguistics which separate out and subcategorise varying approaches to the use of corpus data. But let us first deal with the generalisations. We could reasonably define corpus linguistics as dealing with some set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions. The set of texts or corpus dealt
with is usually of a size which defies analysis by hand and eye alone within any reasonable timeframe. It is the large scale of the data used that explains the use of machine-readable text. Unless we use a computer to read, search and manipulate the data, working with extremely large datasets is not feasible because of the time it would take a human analyst, or team of analysts, to search through the text. It is certainly extremely difficult to search such a large corpus by hand in a way which guarantees no error. The next generalisation follows from this observation: corpora are invariably exploited using tools which allow users to search through them rapidly and reliably. Some of these tools, namely concordancers, allow users to look at words in context. Most such tools also allow the production of frequency data of some description, for example a word frequency list, which lists all words appearing in a corpus and specifies for each word how many times it occurs in that corpus. Concordances and frequency data exemplify respectively the two forms of analysis, namely qualitative and quantitative, that are equally important to corpus linguistics.

The importance of our findings from a corpus, whether quantitative or qualitative, depends on another general factor which applies to all types of corpus linguistics: the corpus data we select to explore a research question must be well matched to that research question. To some extent this is self-evident – a corpus is best used to answer a research question which it is well composed to address. To give an extreme example, there would be little point in exploring the noun classification system of Swahili by looking in a corpus of English newspaper texts. More subtly, we cannot (or can only with some caution) make general claims about the nature of a given language based on a corpus containing only one type of text or a limited number of types of text. Finally, and more subtly still, we must be aware that texts within a corpus that we assume to be homogeneous may, in fact, exhibit differences. For example, a collection of samples from a newspaper, even the same newspaper on the same day, may exhibit entirely predictable differences from one another – the sports section, for example, will draw on different lexis than the international news section. Users of a corpus must be aware of its internal variations, and researchers sometimes use statistical techniques to examine the degree of variability within a given corpus before using it (see Gries 2006c for an example of how to explore such variability within a corpus). The degree of homogeneity of a corpus is then another factor in determining how well matched that corpus is to particular research questions.

We have been discussing the features of texts within a corpus. It should be noted that the term text here denotes a file of machine-readable data. Typically in corpus linguistics these are in fact textual in form, so that each file represents, for instance, a newspaper article or an orthographic transcription of some spoken language. However, the computer files within a corpus do not need to be textual, and there are certainly examples nowadays of files of video data being used as corpus texts, as we will discuss in the next section.

This last point highlights a problem even with the very gross generalisations we have made so far – they are generally accurate, but we can very often find specific
examples that challenge them. For example, although we have said that corpus linguistics always uses machine-readable text, in fact, historically, much work was undertaken on corpora held in paper form; for example Fries (1952) produced a grammar of English based upon such a corpus. Also, while it is true that much research using corpus methods (e.g. McEnery 2005; Davies 2009b; Millar 2009; and many others) uses corpora of millions of words, there are others studies such as those of Ghadessy and Gao (2001) and McEnery and Kifle (2001) which, appropriately, use smaller, specialised corpora that might conceivably have been analysed by hand and eye. Nonetheless, despite the exceptions, the generalisations above characterise much of the work that can reasonably be described as corpus linguistics. Looking beyond these generalisations, research within the field can be divided on the basis of a number of criteria which discriminate quite sharply between types of work. The following features are those which, in our view, most typically distinguish different types of studies in corpus linguistics:

- Mode of communication;
- Corpus-based versus corpus-driven linguistics;
- Data collection regime;
- The use of annotated versus unannotated corpora;
- Total accountability versus data selection;
- Multilingual versus monolingual corpora.

Using these features, we can begin to work out a rough typology of corpus linguistic research, at least in terms of the principles underlying the use of corpora in such studies. Several of the later chapters of this book will be devoted to developing critical overviews of some of the types of corpus linguistics outlined within this typology, including the ‘neo-Firthian’ tradition (Chapter 6) and the variationist tradition (Chapter 5). However, in order to fully understand this typology, we clearly need to define the oppositions above in some detail.

### 1.2 Mode of communication

Corpora may encode language produced in any mode – for example, there are corpora of spoken language and there are corpora of written language. In addition, some video corpora record paralinguistic features such as gesture (Knight et al. 2009), and corpora of sign language have been constructed (Johnston and Schembri 2006; Crasborn 2008).

Corpora representing the written form of a language usually present the smallest technical challenge to construct. Until recently, encoding – and reliably representing on screen – writing systems other than the Roman alphabet was prone to error (Baker et al. 2000). However, with the advent of Unicode (Unicode Consortium 2006), this problem is being consigned to history; Unicode allows computers to reliably store, exchange and display textual material in nearly all of
what is corpus linguistics?

the writing systems of the world, both current and extinct. Written corpora can still be time-consuming and error-prone to produce in cases where the materials have to be either scanned or typed from printed original documents (this is particularly true for handwritten material – see Smith et al. 1998). However, as we will discuss later in this chapter, the increasing availability of a wide range of genres in machine-readable format for most major languages means that the construction of written corpora, except in the context of historical linguistic research, has never been easier.

Material for a spoken corpus, however, is time-consuming to gather and transcribe. Some material may be gathered from sources like the World Wide Web – for example, transcripts of parliamentary debates, called Hansard reports, are produced in the UK. These are readily accessible on the web. Also, Hoffmann (2007a) has gathered transcripts of news broadcasts from the web to represent speech. However, transcripts such as these have not been designed as reliable materials for linguistic exploration of spoken language. Consequently there are ‘serious hazards involved if transcripts that were made by non-linguists for purposes of their own are to be used for linguistic analysis’ (Mollin 2007: 188). Mollin (2007: 208) outlines the dangers of using data such as Hansard, whose transcripts are known to make certain changes to what was actually said:

Some of the changes are due to the fact that Hansard transforms conversation based on the here and now of the situation into a decontextualised report that is also understandable to the distant reader. Adding information on speakers and persons referred to . . . In addition, Hansard omits certain interpersonal and situational references, resulting in a reduction of the very typical parliamentary formulae, e.g. those of turn taking. The picture conveyed to the reader is one where MPs speak orderly one after the other without any apparent meta-comments on how and when to speak.

Given problems such as these, it is hardly surprising that spoken corpus data is more often produced by recording interactions and then transcribing them. Orthographic and/or phonemic transcriptions of spoken materials can be compiled into a corpus of speech which is searchable by computer. These transcriptions may be linked back systematically to the original recording through a process called time-alignment so that, through the computer, it is possible both to easily search a spoken corpus and to hear the portion of the recording that matches a particular search result. This is possible, for example, with the COLT corpus of London teenage speech (Stenström et al. 2002), the International Corpus of English British component (ICE-GB) and the Origins of New Zealand English (ONZE) corpus (Fromont and Hay 2008). The orthographic form of a spoken corpus often normalises the form of the words in the text to standard spellings, meaning that orthographically transcribed material is rarely a reliable source of evidence for research into variation in pronunciation. Phonemically transcribed material is of much more use in this respect, though it tends to be most useful when variant forms can be searched for by reference to a standardised form, typically
1.3 Corpus-based versus corpus-driven linguistics

The difference between corpus-based and corpus-driven language study (to use the terms originally introduced by Tognini-Bonelli 2001) is a topic
that runs through this book. Corpus-based studies typically use corpus data in order to explore a theory or hypothesis, typically one established in the current literature, in order to validate it, refute it or refine it. The definition of corpus linguistics as a *method* underpins this approach to the use of corpus data in linguistics. Corpus-driven linguistics rejects the characterisation of corpus linguistics as a method and claims instead that the corpus *itself* should be the sole source of our hypotheses about language. It is thus claimed that the corpus itself embodies its own theory of language (Tognini-Bonelli 2001: 84–5). This notion of corpus-driven linguistics is closely associated with the work of scholars we will refer to as ‘neo-Firthians’, which will be explored in depth in Chapter 6. In that chapter, we will also revisit and problematise the corpus-based versus corpus-driven distinction. For those who accept it, the corpus-based versus corpus-driven dichotomy creates a basic, binary distinction, under which most works of corpus linguistic research can be sorted into one or the other group. However, our own perspective rejects the notion that the corpus itself has a theoretical status, and thus also rejects the binary distinction between corpus-based and corpus-driven linguistics. From this point of view, all corpus linguistics can justly be described as corpus-based. This point of controversy will be explored in Chapter 6.

### 1.4 Data collection regimes

An important question follows from the observation that corpus studies should match their data to their research question. How can we ensure that the match is good enough? If we want to explore grammatical features in Modern English, we clearly need to match the data we use against the claims we wish to make. To make general claims about spoken English, we would require a suitable spoken dataset. The speech of one person alone is unlikely to provide a suitable basis for such generalisations. So corpus construction, and in particular data collection, emerges as a critical issue for corpus linguistics. Two broad approaches to the issue of choosing what data to collect have emerged: the *monitor corpus* approach (see Sinclair 1991: 24–6), where the corpus continually expands to include more and more texts over time; and the *balanced corpus* or *sample corpus* approach (see Biber 1993 and Leech 2007), where a careful sample corpus, reflecting the language as it exists at a given point in time, is constructed according to a specific sampling frame.

#### 1.4.1 Monitor corpora

The monitor corpus approach, proposed most notably by John Sinclair, seeks to develop a dataset which grows in size over time and which contains a variety of materials. The relative proportions of the different types of materials may vary over time. Monitor corpora could be said to balance any need to be
precise about the composition of a corpus against sheer size – as the corpus grows, we might assume that any skew in the data naturally self-corrects, since there is no consistent skew in the data input. The Bank of English (BoE), developed at the University of Birmingham, is the best-known example of a monitor corpus. The BoE was started in the 1980s (Hunston 2002: 15) and has been continually expanded since that time. At the time of writing, the corpus contains over half a billion words, organised into a general English section (450 million words) and a section containing corpus materials of use in language pedagogy (56 million words). The BoE represents one approach to the monitor corpus; the Corpus of Contemporary American English (COCA; Davies 2009b) represents another. COCA expands over time like a monitor corpus, yet it does so according to a much more explicit design than the BoE. Each extra section added to COCA complies to the same, breakdown of text-varieties. Arguably, this corpus represents something of a halfway house between the sample corpus approach and the monitor corpus approach – a monitor corpus that proceeds according to a sampling frame and regular sampling regime.

While the BoE and COCA are impressive in scale, there is arguably a much larger monitor corpus under construction that covers a wide range of languages and contains a growing record of those languages over time – the World Wide Web.

1.4.2 The Web as Corpus

The concept of Web as Corpus (Kilgarriff and Grefenstette 2003) is very similar in many ways to the idea of the monitor corpus. It takes as its starting point a massive collection of data that is ever-growing, and uses it for the study of language (see, for example, the web-based study of antonyms by Jones et al. 2007, as a good example of the use of the web as a corpus). As well as using standard search engines such as Google to explore the web as a corpus, researchers have also developed interfaces specifically designed to support this use of the web, such as WebCorp (Renouf 2003). The Web as Corpus approach has some specific problems. In contrast to most corpora, the web is a mixture of carefully prepared and edited texts, and what might charitably be termed ‘casually prepared’ material. The content of the web is also not divided by genre – hence the material returned from a web search tends to be an undifferentiated mass, which may require a great deal of processing to sort into meaningful groups of texts. In addition, there is little doubt that the many texts on the web contain errors of all sorts. For example, while writing this book we typed receive and recieve into Google – receive scored 300,000,000 hits, recieve scored 8,670,000 hits. This of course may prove useful – if you wish to investigate common spelling errors, for example. Data like this might also be the basis for a very interesting study in support of spelling reform. However, if this isn’t the sort of thing you are interested in, such errors in the data may well provide unwelcome noise when the analyst approaches the web as a corpus. Given that this kind of noise exists at all
levels of language on the web, it represents a significant issue that the users of the web as a corpus must address. Nonetheless, the web does undoubtedly provide a substantial volume of data which can be selected and prepared to produce corpora suitable for a wide variety of purposes.

By way of illustration, if you wanted to examine the rather loaded phrase *swanning around* in the BNC, you would find only 13 examples of it on which to base your observations. Using Google, we recovered 32,300 examples of texts containing this phrase. Admittedly, those thousands of examples would need to be sorted and sifted before they could be used to explore the phrase reliably. However, there is little doubt that the thousands of examples from Google would allow a more nuanced investigation of this particular phrase than the dozen or so examples in the BNC. So the web is a useful and readily available source of evidence, which can be invaluable in cases where you need a large quantity of data in order to deal with a low frequency of occurrence. However, there is a problem associated with this opportunity: for frequent words or phrases, the number of examples returned by a web search engine may simply be overwhelming, and a good deal of data may have to be discarded. This should be done in accordance with some heuristic which ideally should be applied consistently across all analyses. For example, if we study frequent words like *receive* using the web as a corpus, we may want to study only the first 100 examples that are returned. If we do this for one word in the study, then we should do so for all words in that study, where necessary. Another problem exists with all studies based on web data that is not downloaded and archived appropriately: the web is forever changing. It is difficult to replicate a study done on the web four years ago, for example, as the web will have changed significantly. Given the importance of replicability in experimental procedures (which we will discuss in section 1.6.1), this is an obvious and pressing drawback to the Web as Corpus approach.

### 1.4.3 The sample corpus approach

In contrast to monitor corpora, balanced corpora, also known as *sample corpora*, try to represent a particular type of language over a specific span of time. In doing so, they seek to be balanced and representative within a particular sampling frame which defines the type of language, the population, that we would like to characterise. The population is the notional space within which language is being sampled. So, for example, if we want to look at the language of service interactions in shops in the UK in the late 1990s, the sampling frame is clear – we would only accept data into our corpus which represents service interactions in UK shops in the 1990s. However, if we only collected data gathered in coffee shops, we would not get a balanced set of data for that population – relatively context-specific lexis, such as *latte* and *frappuccino*, would be likely to occur much more frequently than they do in service interactions in general. Phrases which are typical of other kinds of service interactions, such as *Should I wrap that for you?*, might not occur at all. Following the principle of balance, we would seek...
to characterise the range of shops whose language we wanted to sample, and collect data evenly from across that range.

Even if we decided we were only interested in bookshops, coffee shops and supermarkets, we might still wish to ensure that the shops sampled from were in some sense typical, and that we gathered data from them in such a way as to avoid introducing skew into our dataset. So, we might care to ensure that we did not sample from bookshops which sell only antiquarian books, if we were concerned that the interactions there could be atypical of bookshops in general. Similarly, we might want to ensure that the proportions of data in our corpus reflect, in some way, the numbers of each type of interaction of interest that actually occur. If we had 90 per cent of our data from bookshops, 8 per cent from coffee shops and 2 per cent from supermarkets, when we know that there are a hundred supermarkets for every bookshop, we might well feel that our corpus design was less than ideal. We would have to choose the locations to sample from, and the relative proportions of different types of data to collect, with the aim of achieving representativeness for the data in a corpus. Of course, this simple example presents only one approach to representativeness; see Leech (2007) for a critical exploration of this concept.

Corpora which seek balance and representativeness within a given sampling frame are snapshot corpora. A good example of a snapshot corpus is the Lancaster-Oslo/Bergen (LOB) corpus. This represents a ‘snapshot’ of the standard written form of modern British English in the early 1960s. Table 1.1 gives the sampling frame within which the data for the LOB corpus was gathered.

For each category, samples of data were gathered, with each sample being of roughly similar length (2,000 words). The samples were taken from a variety of sources within each broad sampling domain. The resulting corpus is 1 million words in size. The LOB corpus demonstrates how a snapshot corpus, used in concert with corpora constructed using the same sampling frame, can allow us to undertake a wide range of contrasts and comparisons. The same sampling frame used for LOB has also been used to collect corpora of written British English at spaced intervals (mostly of thirty years) through the twentieth and early twenty-first centuries. This allows the effects of diachronic change to be studied in this variety of English (see Leech 2004; Baker 2009; Leech et al. 2009). This approach to exploring diachronic change is analogous to stop-motion photography – slow-moving changes become visible when a snapshot is taken at discontinuous intervals. It is also possible to study diachronic change with a large monitor corpus, though different techniques may be needed in order to capture slow-moving change over time. Using snapshot corpora, we can also look at synchronic differences in varieties of English. The LOB sampling frame was adopted from one originally developed to construct a corpus of written American English from 1961, the Brown Corpus (Francis and Kučera 1964; Kučera and Francis 1967) – so comparing LOB and Brown, we can investigate differences in the two language varieties while controlling for sampling and the effects of diachronic change. We will return to the study of synchronic and diachronic variation using the LOB and Brown corpora in section 5.3.
WHAT IS CORPUS LINGUISTICS?

Table 1.1 The LOB Corpus Sampling Frame (after Hofland and Johansson 1982: 2)

<table>
<thead>
<tr>
<th>Category mnemonic</th>
<th>Description</th>
<th>Number of text samples in this category</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Press: reportage</td>
<td>44</td>
</tr>
<tr>
<td>B</td>
<td>Press: editorial</td>
<td>27</td>
</tr>
<tr>
<td>C</td>
<td>Press: reviews</td>
<td>17</td>
</tr>
<tr>
<td>D</td>
<td>Religion</td>
<td>17</td>
</tr>
<tr>
<td>E</td>
<td>Skills, trades and hobbies</td>
<td>38</td>
</tr>
<tr>
<td>F</td>
<td>Popular lore</td>
<td>44</td>
</tr>
<tr>
<td>G</td>
<td>Belles lettres, biography, essays</td>
<td>77</td>
</tr>
<tr>
<td>H</td>
<td>Miscellaneous (government documents, foundation reports, industry reports, college, catalogue, industry house organ)</td>
<td>30</td>
</tr>
<tr>
<td>J</td>
<td>Learned and scientific writings</td>
<td>80</td>
</tr>
<tr>
<td>K</td>
<td>General fiction</td>
<td>29</td>
</tr>
<tr>
<td>L</td>
<td>Mystery and detective fiction</td>
<td>24</td>
</tr>
<tr>
<td>M</td>
<td>Science fiction</td>
<td>6</td>
</tr>
<tr>
<td>N</td>
<td>Adventure and western fiction</td>
<td>29</td>
</tr>
<tr>
<td>P</td>
<td>Romance and love story</td>
<td>29</td>
</tr>
<tr>
<td>R</td>
<td>Humour</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>500</td>
</tr>
</tbody>
</table>

1.4.4 Balance, representativeness and comparability

Balance, representativeness and comparability are ideals which corpus builders strive for but rarely, if ever, attain. In truth, the measures of balance and representativeness are matters of degree. Váradi (2001) has been critical of the failure of corpus linguists to fully define and realise a balanced and representative corpus. Even proposals, such as those of Biber (1993), to produce empirically determined representative corpora have not actually been pursued. Biber’s proposal for representativeness to be realised by measuring internal variation within a corpus – i.e. a corpus is representative if it fully captures the variability of a language – has yet to be adopted in practice. It is also only one of many potential definitions of representativeness, as Leech (2007) points out. However, though balance and representativeness remain largely heuristic notions, decided on the basis of the judgement of linguists when they are building a corpus, this does not mean to say that the concepts are of no value. Similarly, while some corpora designed to be comparable to each other can clearly make a claim for balance and representativeness, others may only do so to a degree. Leech (2007: 141–3) usefully summarises a series of problems encountered in building comparable corpora of British English to explore diachronic variation: notably,