Cambridge Textbooks in Linguistics

Corpus Linguistics

Tony McEnery and Andrew Hardie

CAMBRIDGE

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.[1] 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).[2] 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

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.[3] 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)[4] 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

the orthographic transcription. In this way, the differing phonemic transcriptions corresponding to a single standardised form in different contexts can be compared and contrasted (as is possible, for example, in the Spoken English Corpus; Knowles *et al.* 1996). An interesting issue arises when compiling or analysing a spoken corpus of a language for which there is no written form, or where the written form is not easily rendered in machine-readable form. In this case it may be necessary to rely on phonemic transcription alone, or to decide upon an orthographic transcription scheme that allows for recovery of forms which are equivalent yet vary phonetically.

Corpora which include gesture, either as the primary channel for language (as in sign language corpora) or as a means of communication parallel to speech, are relatively new. The integration of video analysis with textual analysis is clearly crucial for the development and use of such corpora. This may be achieved by using software packages, such as Eudico (Brugman *et al.* 2002) and the Digital Replay System,[5] which allow text, sound and video to be brought into alignment with one another for the purposes of searching and analysing the data. Given the advances in technology that were required in order to handle such data, it is unsurprising that corpus linguistic studies focusing on the visual medium are only just beginning to be undertaken on a truly large scale, for example investigating the relationship between gesture and speech (Carter and Adolphs 2008), or constructing large corpora of sign language material (Johnston and Schembri 2006). Novel analytical schemes have also needed to be developed to begin the process of analysing the video streams that form the raw data of these types of corpora (for examples see Wittenburg *et al.* 2002 and Knight *et al.* 2009).

These basic distinctions in mode of communication do not map simply onto corpus data – many corpora contain data from more than one mode, such as the British National Corpus (BNC; Aston and Burnard 1998), which contains both speech and writing. However, the medium of communication itself does produce a distinction which is linguistically meaningful. Large-scale work contrasting spoken and written language has led to a much deeper appreciation of how remarkably different the two can be, as is shown at the level of grammar, for example, by Biber *et al.* (1999) and Carter and McCarthy (1995). The differences are such that some linguists, notably Brazil (1995), have made the claim that the grammar of speech and that of writing are not merely distinct but entirely different (see section 4.6). Consequently, thinking about corpora in terms of mode of production is not just a matter of different data collection and technical issues; we would argue that it is, rather, linguistically a very real distinction.

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.[6] 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 *frapuccino*, 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.

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

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

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,

problems relating to the evolution over time of the genres that are balanced in those corpora. The changing nature of genre makes claims of comparability when looking at diachronic variation much more tendentious than similar claims for the synchronic Brown/LOB comparison, for example. As Leech (2007: 143–4) notes, the debate around balance, representativeness and comparability might lead researchers:

> to reject these concepts as being ill-defined, problematic and unattainable. My attitude is different from this . . . these are important considerations, and even if we cannot achieve them 100 per cent, we should not abandon the attempt to define and achieve them. We should aim at a gradual approximation to these goals, as crucial desiderata of corpus design. It is best to recognise that these goals are not an all-or-nothing: there is a scale of representativity, of balancedness, of comparability. We should seek to define realistically attainable positions on these scales, rather than abandon them altogether.

There is little doubt that, as the corpus approach to language develops, the concepts of balance and representativeness will undergo further critical scrutiny. This in turn should lead to incrementally better definitions of these terms.

1.4.5 'Opportunistic' corpora and minority and endangered languages

The monitor versus snapshot corpus distinction provides us with a ready framework for categorising corpora which make some claim to represent a particular language in general. However, it also must be noted that there are many collections of data, reasonably described as corpora, which do not necessarily match the description of either a monitor or a snapshot corpus comfortably. Such corpora are best described as *opportunistic* corpora. These corpora make no pretension to adhere to a rigorous sampling frame, nor do they aspire to deal with issues of skew by the collection of an ever-larger body of data, as monitor corpora may. Rather, they represent nothing more nor less than the data that it was possible to gather for a specific task. Sometimes technical restrictions prevent the collection of large volumes of data to populate some idealised sampling frame. This was particularly true prior to the widespread introduction of electronic publishing and the general availability of electronic text in a range of languages on the web. Some early corpora were not built along principled lines according to the demands of a specific research question; rather, they were constructed using whatever relevant material could be accessed in electronic form. Corpora such as the American Printing House for the Blind corpus (Black *et al.* 1993) and the Hansard Corpus (Berger *et al.* 1994) were built in order to exploit materials from what were, at the time, two of the very few text producers who created machine-readable versions of texts. This problem clearly no longer generally applies to English or most other major languages, but it still persists for some languages.

It is likely that for languages with a written form, more and more machine-readable textual material will become available over time, allowing them to be readily studied. Consider the general division of languages into four broad types suggested by McEnery and Ostler (2000):

1. Official majority languages (e.g. English in the UK, Portuguese in Portugal).
2. Official minority languages (e.g. Welsh in the UK).
3. Unofficial languages (both large, e.g. Kurdish in Turkey, and relatively small, e.g. Sylheti in the UK).
4. Endangered languages (e.g. Guugu Yimidhirr in Australia).

It is fair to say that types 1 and 2 are better supplied with corpus data than 3 and 4 for a range of non-linguistic reasons. Official languages typically have governments with money associated with them. These governments typically publish material in the official language, often on the web. They also, at times, fund corpus-building projects. Unofficial languages suffer from a lack of official recognition, and hence state funding. Furthermore, if the language is associated with an oppressed group, the language itself may be suppressed. The issue with endangered languages is obvious – very few speakers producing little material relative to the larger languages. It may also be the case that endangered languages are also suppressed, making their situation yet worse.

A significant problem arises in the context of analysts approaching spoken data in particular: converting spoken recordings into machine-readable transcriptions is a very time-consuming task. This in itself means that, without significant financial support or plenty of available time, some analysts choose to work on small datasets when much larger datasets would arguably be more appropriate for their task. Analysts may feel, rightly given the resources available, that working with a small sample may be sufficient for their purposes, and that while a larger dataset might yield slightly different results, they face the prospect of 'a huge amount of work and planning for very small returns' (Holmes 1996: 168). A researcher must at times be guided by pragmatism.

Finally, even with a huge amount of work and planning, it may simply be impossible to build an ideal corpus for a given language – if the language is dead or dying and the material to construct a large, balanced corpus is not available and simply never will be. To consider an extreme example, the Indus Valley civilisations based around Harappa and Mohenjo-daro flourished between approximately 2,500 and 1,900 BCE. The total stock of written material that remains to represent the language used by that civilisation consists of 3,700 inscribed objects (Robinson 2009: 268). It is unlikely that future archaeological digs will significantly alter the extent of this stock of text. If we want to build a corpus from these objects, perhaps to try to decode this as yet undeciphered script, the amount of material to draw on is quite finite – the language is dead and the writing system is no longer used. In all likelihood we have the great majority of surviving 'texts' in our possession already. No native speaker of the

language of the Indus Valley will ever again exist to produce more texts using this writing system. When dealing with an extinct language for which a greater body of literature survives, such as Classical Latin, Gothic or Old English, our situation is different in degree but not in kind from the Indus Valley case: our only choice in building a corpus is to select some or all of the texts that have made it through the centuries.

In summary, while the notions of monitor and snapshot corpora provide us with relatively idealised models of corpus construction, it should be noted, and accepted, that the corpora that we use and construct must sometimes be determined by pragmatic considerations.

1.5 Annotated versus unannotated corpora

A further way in which studies in corpus linguistics vary relates to whether or not linguistic analyses are encoded in the corpus data itself. Such encoding, called *corpus annotation*, may be achieved either by editing the data to include within it some analysis, or by having the analysis stored separately but linked in to the data. For example, we may wish to annotate a corpus to show parts of speech, assigning to each word the grammatical category we claim it has in its context. So, for example, when we see the word *talk* in the sentence *I heard John's talk and it was the same old thing*, we would assign it the category 'noun' in that context. In doing so, we might edit the text directly, assigning some mnemonic code (such as N) to make it clear that in this case the word is a noun. In a simple case, we may just attach the mnemonic code to the word in question with an underscore – *talk_N*.[7] Rather than edit the text directly, however, it is also possible to store annotations like this separately from the data itself, using computer programs to combine, integrate and disentangle the text and annotations as the analyst desires. This so-called 'stand-off' annotation is preferred by some analysts (e.g. Thompson and McKelvie 1997). However, given a systematic encoding of annotations directly in a corpus, it is a trivial matter to remove them if desired, so the arguments in favour of stand-off annotation seem to boil down more to a question of methodical neatness or elegance rather than denoting anything fundamental in nature.[8]

While the phrase *corpus annotation* may be unfamiliar to some linguists, the basic operation it describes is not – it is directly analogous to the analyses of data that have been done using hand, eye and pen for decades. Corpus annotation is, then, a commonplace of linguistics. If it varies from usual practice at all, it is in the scale on which it is applied. In Chomsky (1965), twenty-four invented sentences are analysed; in the parsed version of LOB, a million words are annotated with parse trees. Nonetheless, it is important to note that, setting scale aside, corpus annotation is largely the process of providing – in a systematic and accessible form – those analyses which a linguist would, in all likelihood, carry out anyway on whatever data they worked with.

On the basis of this somewhat brief description of corpus annotation, a reader would be forgiven for thinking that the distinction between annotated and unannotated corpora is based simply on whether or not the corpus has been analysed in a particular way *yet*. Those corpora which have already been analysed in some way are annotated, those which have yet to be analysed are not. This distinction in itself, however, is so trivial that it would hardly constitute a major dimension along which research in corpus linguistics can vary. What makes this dimension salient is the fact that some linguists object to annotation – either per se, or when undertaken manually rather than automatically by a computer. Opposition to annotation is typically associated with neo-Firthian corpus linguistics and the corpus-driven approach, as will be discussed in Chapter 6. However, in brief, arguments against annotation are largely predicated upon the purity of the corpus texts themselves, with the analyses being viewed as a form of impurity. This is because they impose an analysis on the users of the data, but also because the annotations themselves may be inaccurate or inconsistent (Sinclair 1992). Such claims are interesting because, as has been noted, corpus annotation is the manifestation within the sphere of corpus linguistics of processes of analysis that are common in most areas of linguistics. To identify problems with accuracy and consistency in corpus annotation is, in principle at least, to identify flaws with analytical procedures across the whole of linguistics. It is because of the issues of accuracy and consistency, in particular, that some linguists prefer to use unannotated corpora. But this does not mean to say that such linguists do not analyse the data they use; rather, it means that they leave no systematic record of either their analysis or their errors which can easily and readily be tied back to the corpus data itself.

1.6 Total accountability versus data selection

So far we have focused on ways in which corpora vary in their design. Corpora may also vary, however, in how they are used by the analysts who exploit them. A key difference here is the contrast between *total accountability* and *data selection*.

1.6.1 Total accountability, falsifiability and replicability

It has been argued that a significant advantage of using corpora is that corpora allow analysts to approach the study of language within the context of the scientific method (Leech 1992). A core principle of Leech's approach within this framework is total accountability (Leech 1992: 112). If you approach a corpus with a specific theory in mind, it can be easy to unintentionally focus on and pull out only the examples from the corpus that support the theory (this is technically called a *confirmation bias*). But the theory can never be shown to be false by such

an approach, even in principle. As such, this approach runs counter to one of the key features of the scientific method identified by Popper ([1934] 2006: 18), namely *falsifiability*. The principle of total accountability is, simply, that we *must not* select a favourable subset of the data in this way. When approaching the corpus with a hypothesis, one way of satisfying falsifiability is to use the entire corpus – and all relevant evidence emerging from analysis of the corpus – to test the hypothesis. This principle is the reason for the quantitative nature of many corpus-based methods. Minimally, however, where there is too much evidence for using the entire corpus to be practical, the analyst must at least, as Leech suggests, avoid conscious selection of data. Short of using the corpus in its totality, total accountability can in principle be preserved by using an unbiased (e.g. randomised) subsample of the examples in the corpus. If it were permissible, in corpus research, to filter out or ignore examples or statistics from the corpus that do not fit the hypothesis under investigation, then the corpus could support such a bewildering variety of potentially contradictory hypotheses that the use of corpus data would be fatally undermined. To put it simply, there should be no motivated selection of examples to favour those examples that fit the hypothesis, and no screening out of inconvenient examples. Such a statement represents an ideal for the use of corpus data that most would find difficult to challenge.

However, there is a criticism to be levelled at such an approach: the corpus *itself* is necessarily a finite subset of a much larger (and in principle non-finite) entity, language. So the corpus itself represents a selection and screening of data. Therefore, any claim of total accountability in corpus linguistics must be moderated. We can only seek total accountability relative to the dataset that we are using, not to the entirety of language itself. This criticism is not, of course, unique to linguistics. An obvious parallel is astronomy, where astronomers theorise on the basis of the subset of the Universe that is visible to them. They expand their dataset over time, and each generation of astronomers seeks to falsify the findings of the previous generations of astronomers as they push forward the boundaries of the field. A very similar model is developing in linguistics, now that it has become possible to expose linguistic theories to testing by large-scale observation. Based on this analogy, we can say that, like an astronomer, a corpus linguist can work in accordance with the scientific method, and produce potentially falsifiable results, while not being totally accountable in the strictest sense.

But moderating the claim of total accountability in the light of the finite size of the corpus does raise one troubling possibility. An analyst may, by chance or design, construct a dataset that misrepresents the language such that the analysis of this dataset supports a faulty theory. While we must be mindful of this possibility, an analogy with astronomy may help once again. Let us imagine an astronomer, at some point in the past, seeking to develop a model of moons based on data from the Earth, Mars and Jupiter. They then conclude, from that dataset, that all planets have moons. The problem here is with the dataset – it has unwittingly been drawn from a set of planets which happen to have moons. If Mercury or Venus, which lack moons, had been in the dataset, the conclusion would have been

different. The answer to the problem in astronomy is the same as in linguistics, and emerges from another key feature of the scientific method: *replicability*. A result is considered replicable if a reapplication of the methods that led to it consistently produces the same result. This process of checking and rechecking may be done with the same dataset or it may be done with new datasets. In Popper's theory, falsifiability is of higher priority than replicability as a key to verification in the scientific method. The ability to replicate a result, whether experimental or observational, is, nonetheless, still clearly central to scientific practice. In all the sciences, new results are typically considered provisional until they are known to be replicable – and in many cases, it is precisely through that process of continuous checking of results as theories develop and expand that falsifiability is achieved.

Like the natural sciences, corpus linguistics has in many cases appealed to the notion of the replicable result for credibility (see Doyle 2005 for a good critical overview of the engagement of corpus linguistics with replicability). In particular, replicability helps us address the problem of the limited dataset outlined above. Attempts to replicate the astronomical result that all planets have moons will, eventually, find that in a wider dataset of planets, the rule does not hold. Similarly, an incorrect or incomplete result that stems from the finite size of a corpus is likely to be found out when corpus linguists recheck that result against other datasets. So as long as this process of checking and replication runs its course, and given sufficient time and data, bias in the data of the sort we have outlined is routinely discovered and removed. There is evidence of this happening already in linguistics in general, and corpus linguistics in particular. A good example of work undertaken on one corpus being revised when further corpus data became available is Leech's (1971, 2004a) work on non-finite verbs (see also section 2.2). In sum, then, total accountability to the data at hand ensures that our claims meet the standard of falsifiability; total accountability to *other* data in the process of checking and rechecking ensures that they meet the standard of replicability; and the combination of falsifiability and replication can make us increasingly confident in the validity of corpus linguistics as an empirical, scientific enterprise.

1.6.2 Data selection – not (necessarily) a bad thing

Given what has been said about total accountability, you may wonder that analysts would ever approach a corpus seeking a single example, or a subset of carefully selected examples. Not only do some analysts do just that, in certain circumstances it may actually be the right thing to do. Indeed, in an important sense, approaching a corpus in search of a specific type of result may be entirely in line with the scientific method. Most importantly, we may seek in a corpus a specific example which, in itself, falsifies a hypothesis – thereby making the totality of the data in some sense irrelevant. One example alone may be enough to falsify a claim. In a corpus of a million sentences, the one sentence that

does not conform to a hypothesis is the only sentence that really matters for considering the hypothesis in question. This may be illustrated by returning to our astronomy parallel. Given the hypothesis that all planets have moons, if we have data available from a thousand planets, the fact that 999 of them have moons is not as important – from the point of view of defending the hypothesis – as the fact that one planet has no moons at all. Likewise, if the hypothesis we are looking at is that some particular linguistic form never occurs, then the only part of the corpus that is really relevant is the part where that linguistic form *does* occur, thus falsifying the hypothesis. To put this in general terms, a single example may falsify a hypothesis, leading to the revision, or abandonment, of that specific hypothesis. In that sense, approaching a corpus to find a single example is entirely consistent with both the scientific method and with the principle of total accountability.

A more contentious manifestation of utilising only selected parts of a corpus arises when researchers use the corpus simply as a bank of examples to illustrate a theory they are developing – this is sometimes called *corpus-informed* research. This clearly does run counter to the scientific method, insofar as there is no attempt to account for the rest of the (potentially falsifying) evidence in the corpus. However, some researchers have articulated an interesting motivation for using corpora in such a fashion. The premise is not unlike that which drives corpus linguistics to validate and revalidate hypotheses – namely, that the corpus is finite, but language is not. Some researchers argue that corpora, while a helpful guide or source of examples, cannot give sufficient access to language to the extent that so-called 'qualitative' approaches to the data should be abandoned. A good example of this has emerged in Critical Discourse Analysis (CDA).

CDA has traditionally been approached by the detailed analysis of single texts or small numbers of texts. On the basis of that detailed analysis, general claims about the use of language in society have then been made. Over time, as evidence from the analysis of individual texts has accumulated, overarching theories of how discourses work in society have emerged; and generic claims about the structure and nature of such discourse, focused, for example, on specific words or classes of word such as pronouns, have been made. These general observations, based on a small number of texts, have been exploited within an overarching theoretical framework based upon some theory of power relations. Since the mid-1990s, attempts have been made to integrate the general methodological approach of corpus linguistics with CDA by researchers such as Mautner (see Hardt-Mautner 1995, 2000; Mautner 2009), Koller and Mautner (2004), O'Halloran and Coffin (2004), Baker (2004, 2006, 2009) and Orpin (2005). A general issue with most of these attempts at integration has been one of balance – studies have tended either to focus mainly on either corpus linguistics or CDA at the expense of the other. Corpus-based studies may have explored discourse and its relation to power, but they have typically not been explicitly informed by CDA theory and its traditional methods, or else they have not aimed to contribute to a particular discourse-oriented theory (e.g. Stubbs 1994; Krishnamurthy 1996). Similarly,

CDA researchers have at times used data and techniques which are undoubtedly inspired by work in corpus linguistics, but have not sought to engage fully with the corpus approach (e.g. Fairclough 2000; Kovàcs and Wodak 2003). Research which is principally CDA-oriented tends to make limited or casual use of a corpus or corpus-based techniques. Sometimes, the corpus is used simply as a repository of examples (e.g. Flowerdew 1997) and no effort is made to apply the principle of total accountability that is generally accepted within corpus linguistics. Also, CDA studies making use of corpora have in general tended to avoid carrying out quantitative analyses beyond the simplest of descriptive statistics (see also Stubbs 1997: 104), preferring to undertake qualitative analyses using concordances.

Why do some researchers in CDA only engage minimally with corpus data? An important argument presented by such researchers relates to the depth of analysis that they want using the data they have – they wish to undertake a detailed analysis of a small amount of data, taking into account not just the text itself, but also the social context in which it was produced and the social context in which it was interpreted. This work is so labour-intensive that a large-scale study using the corpus may not be possible.[9] This argument has some weight. However, there is also the possibility of striking a balance where the corpus data itself is used in the framework of total accountability, but the detailed analysis is reserved for a subset of the data, once those hypotheses that are testable in practical terms on the whole corpus have been tested (KhosraviNik 2009). Nonetheless, it is still the case that many researchers prefer to work with small amounts of data in detail rather than engage with large corpora.

1.7 Monolingual versus multilingual corpora

Another obvious way in which corpora vary relates to the number of languages represented in the corpus.[10] Many corpora are monolingual in the sense that, while they may represent a range of varieties and genres of a particular language, they are nonetheless limited to that one language. So the International Corpus of English (ICE; see also section 4.2), for example, is a large monolingual corpus – it represents one language, English, though it allows linguists to compare and contrast a number of international varieties of that language. Monolingualism in corpora may be a matter of degree rather than an absolute. The BNC, for example, does contain some foreign words and speech produced by non-native English speakers (Aston and Burnard 1998: 127). However, the appearance of such data in the BNC does not reflect its primary purpose, which is to represent modern British English. The fact that some material in a language other than English was inadvertently collected does not mean that we should regard this corpus as anything other than what it claims to be – a monolingual corpus of English. However, the BNC could conceivably be considered (part of) a multilingual corpus if it were brought together with

a range of other corpora, of comparable size, scale and sampling frame, which happen to represent languages other than English. In order to understand this point, we need to consider the variety of multilingual corpora available.

When we refer to a corpus involving more than one language as a multilingual corpus, we are using the term *multilingual* in a broad sense to indicate 'two or more languages'; in a narrower sense, a multilingual corpus must involve at least three languages, while those involving only two languages are conventionally referred to as *bilingual* corpora. Given that corpora involving more than one language are a relatively new phenomenon, with most research hailing from the early 1990s (e.g. the English-Norwegian Parallel Corpus or ENPC; see Johansson and Hofland 1994), it is unsurprising to discover that there is some confusion surrounding the terminology used in relation to these corpora. Generally, there are three types of corpora involving more than one language:

- Type A: Source texts in one language plus translations into one or more other languages, e.g. the Canadian Hansard (Brown *et al.* 1991), CRATER (McEnery and Oakes 1995; McEnery *et al.* 1997).
- Type B: Pairs or groups of monolingual corpora designed using the same sampling frame, e.g. the Aarhus corpus of contract law (Faber and Lauridsen 1991), the Lancaster Corpus of Mandarin Chinese (McEnery *et al.* 2003), which uses the same sampling frame as LOB and Brown.
- Type C: A combination of A and B, e.g. the ENPC (Johansson and Hofland 1994), the EMILLE corpora (Baker *et al.* 2004).[11]

Different terms have been used to describe these types of corpora. For Aijmer *et al.* (1996) and Granger (1996: 38), type A is a *translation corpus* whereas type B is a *parallel corpus*; for Baker (1993: 248; 1995, 1999), McEnery and Wilson (2001: 70) and Hunston (2002: 15), type A is a *parallel corpus* whereas type B is a *comparable corpus*; and for Johansson and Hofland (1994) and Johansson (1998: 4–5), the term *parallel corpus* applies to both types – A and B. Barlow (1995, 2000: 110) certainly interpreted a 'parallel' corpus as type A when he developed the *ParaConc* corpus tool. It is clear that some confusion centres around the term *parallel*.

When we define different types of multilingual corpora, we can use different criteria, for example the number of languages involved and the content or the form of the corpus. But when a criterion is decided upon, the same criterion must be used consistently. For example, we can say a corpus is monolingual, bilingual or multilingual if we take the number of languages involved as the criterion for definition. We can also say a corpus is a translation (L2) or a non-translation (L1) corpus – type A or type B in the framework above – if the criterion of corpus content is used. But if we choose to define corpus types by the criterion of corpus form, we must use it consistently. Then we can say a corpus is parallel if the corpus contains source texts and translations in parallel, or it is a comparable corpus if its subcorpora are comparable by applying the same sampling frame. It

is illogical, however, to refer to corpora of type A as 'translation' corpora by the criterion of content while referring to corpora of type B as 'comparable' corpora by the criterion of form. Consequently, in this book, we will follow Baker's terminology in referring to type A as parallel corpora and type B as comparable corpora. As type C is a mixture of the two, corpora of this type should be referred to as comparable corpora in a strict sense.

A comparable corpus can thus be defined as a corpus containing components that are collected using the same sampling method, e.g. the *same proportions* of the texts of the *same genres* in the *same domains* in a range of *different languages* in the *same sampling period*. We previously observed that the BNC could conceivably become a sub-part of a comparable corpus if corpora similar to the BNC were collected in a range of languages. The resulting collection of corpora could be viewed as a multilingual corpus. However, the sub-parts of this multilingual corpus could also be considered monolingual corpora in their own right. Where there is an equivalence of sampling frames between corpora in different languages, they may be viewed and used as either monolingual or multilingual corpora as necessary. The subcorpora of a comparable corpus are not translations of each other. Rather, their comparability lies in the similarity of their sampling frames.

By contrast, a parallel corpus can most easily be defined as a corpus that contains native language (L1) source texts and their (L2) translations. This definition assumes that parallel corpora are unidirectional (e.g. from English into Chinese or from Chinese into English, but not both). This is currently the most common form of parallel corpus; for instance, the CRATER and EMILLE corpora already mentioned, as well as MULTEXT and P-ACTRES (Izquierdo *et al.* 2008), are unidirectional. However, there are some parallel bidirectional corpora, such as the Portuguese/English COMPARA corpus (Frankenberg-Garcia and Santos 2003),[12] the Nepali/English parallel section of the Nepali National Corpus (Yadava *et al.* 2008) and the English Swedish Parallel Corpus (Altenberg and Aijmer 2000); and there also exist multidirectional corpora (see, for example, the ECC-TEC corpus, Laviosa 2002). Arguably texts which are produced simultaneously in different languages (e.g. EU and UN regulations) can also be classed as parallel data (Hunston 2002: 15).

While parallel and comparable corpora are supposed to be used for different purposes (typically translation research and contrastive studies respectively; see Johansson 2007), the two are also designed with different focuses. For a comparable corpus, the sampling frame is essential. All the components must match with each other in terms of what types of texts they sample, in what proportions, from what periods. For the translated texts in a parallel corpus, the sampling frame is irrelevant, because all of the corpus components are exact translations of each other. Once the source texts have been selected in the first place, there is no need to worry about the sampling frame in the other language. However, this does not mean that the construction of parallel corpora is easier. For a parallel corpus to be useful, an essential step is to *align* the source texts and their

translations, annotating the correspondences between the two at the sentence or word level (see Oakes and McEnery 2000 for an overview). While this would ideally be accomplished using a computer program rather than manual analysis, the automatic alignment of parallel corpora is not a trivial task for some language pairs (Piao 2000, 2002).

1.8 Summary

By looking at a series of defining features in corpus linguistics, this chapter has explored ways in which the construction and use of corpora vary. In doing so, we have highlighted some of the differences that exist between linguists in their use – and basic conception – of corpus linguistics. In the two chapters that follow, we will shift the focus of our discussion to consider a range of more practical matters that corpus linguists face – how to annotate corpus data, how to analyse it and how to employ statistical techniques. We will also consider some of the constraints placed on corpus research by legal and ethical considerations. Throughout this discussion, however, key themes of this chapter will be returned to as they impact upon these practical issues. For example, the decision on whether to annotate or not is an important issue of principle as well as being a practical consideration. Likewise, the World Wide Web presents analysts with specific legal challenges, and the collection of spontaneous speech may bring with it significant ethical issues. So this chapter has raised some core issues which will resurface in a number of ways not only in the two chapters that follow, but also throughout the rest of this book.

Further reading

There is a growing body of books that deal in general with the topic of corpus linguistics. For those readers particularly interested in an approach to corpus linguistics which focuses upon genre analysis and textual variation, Biber *et al.* (1998) is both comprehensive and strongly recommended. With a somewhat different focus, Kennedy (1998) covers in some detail how corpus linguistics and language teaching in particular have intersected. For a general overview of corpus linguistics, with a discussion of its fall from favour in the mid-twentieth century, McEnery and Wilson (2001, see especially Chapter 1) should provide a rewarding read.

While these texts do contain some practical advice, other introductions to corpus analysis have a more hands-on focus. McEnery *et al.* (2006) is the only book that we are aware of which provides a 'how-to' approach to using a wide range of corpus search software. By contrast, Hoffmann *et al.* (2008) build their introduction to corpus linguistics around a single tool, BNCweb. Adolphs (2006) has yet a different emphasis, considering the analysis of *texts* as well as corpora

via the methods of corpus linguistics. Finally, Anderson and Corbett (2009) present an introduction to corpus methods using a range of online analysis tools, a kind of software which we will discuss in detail in section 2.5.4.

Of general interest are the various handbooks and readers for corpus linguistics that have been published. Lüdeling and Kytö (2008) and O'Keefe and McCarthy (2010) are two recent handbooks containing a very wide range of helpful readings in corpus linguistics. Both are, however, somewhat expensive and are probably best sought out via a library. More accessible in price is the reader edited by Sampson and McCarthy (2004). This contains a series of 'classic' papers covering a wide range of topics in corpus linguistics.

For those readers interested in the monitor corpus approach, Sinclair (1991), while now somewhat difficult to buy, is available in many libraries. It is a concise introduction not only to the ideas underlying the monitor corpus, but also to many of Sinclair's other thoughts on language. For some reading suggestions on the Web as Corpus approach specifically, please see the further readings section in Chapter 3.

It is harder to make suggestions for readings on non-English corpus linguistics. While there is an increasing amount of research using corpora of other languages, the main textbooks in the field generally remain engaged with English. For this reason, the primary literature – as found in edited collections such as Johansson (2007) and journals such as *Corpora*, *Corpus Linguistics and Linguistic Theory*, and the *International Journal of Corpus Linguistics* – currently represents the best source of material related to non-English corpus linguistics.

Practical activities

As explained in the foreword, we have designed the exercises in this book to be completed with *any* concordancer and with whatever corpus data you have available. The practical exercises for Chapter 1 are a set of very general tasks that should help you find your way around your concordancer if you are not entirely familiar with it. Either using the 'help' file of the software, or else simply by trial and error, try to find out the following things about your concordancer – all of which you will need to know for exercises later in this book.

(A1-1) Firstly, investigate the basic set-up procedures of your software.

- How do you load a corpus into your concordance tool?
- How do you change to a different corpus?
- Does the entire corpus have to be in a single text file, or can your concordancer handle a corpus consisting of many files?
- Does your concordancer need the texts to be in a particular format, or is simple plain text OK?

(A1-2) Next, look at how the concordancing function works.

- How do you search for a particular word?

- Can you search for annotations such as part-of-speech tags, lemmata or semantic tags – assuming, of course, that they are present in your corpus?
- Are searches case-sensitive (treat <A> and <a> differently) or case-insensitive (treat them the same)? Can you change this behaviour?
- Can you *thin* concordances, i.e. reduce the number of results that are displayed?
- How do you save or export a concordance for later reference?

(A1-3) Finally, work out what the statistical capabilities of your concordancer are.

- How can you get a frequency list (of words or tags) in your concordancer?
- Can you get basic corpus summary statistics – such as total number of words (tokens), type–token ratio and so on?
- Can you produce tables of collocation statistics from a concordance?
- Is there a keywords function? If so, how does it work? Can it be adjusted to analyse key tags?
- Can you get a frequency list of *n-grams* (also known as *clusters* or *multi-word units*)?
- How do you save or export these statistical results?

Questions for discussion

(Q1-1) Look at the breakdown of genres within the (hypothetical, non-existent!) corpus of modern British English described in Table 1.2. Is it balanced? Is it representative? Can these claims be made for any corpus sampling frame in an absolute sense, or must they always be qualified?

Table 1.2 *A hypothetical corpus*

Type of text	Number of words
Press (news reports)	7,500,000
Press (opinion columns)	5,000,000
Press (sports news)	5,000,000
Press (culture news and reviews)	5,000,000
Published fiction (books and short stories)	3,500,000
Unpublished fiction (gathered from the Internet)	1,500,000
General non-fiction books	4,000,000
Academic journals (humanities)	500,000
Academic journals (sciences)	500,000
Television programme transcripts (talk shows)	750,000
Television programme transcripts (news broadcasts)	750,000

(Q1-2) Have a look at three or four research papers from the recent primary literature on corpus linguistics – if you can't think what to look at, we suggest any of the following: Culpeper (2009), Calude (2008), Chung (2008), Diani (2008), Hunston (2007), Oakes and Farrow (2007), Inaki and Okita (2006), Biber and Jones (2005), McIntyre *et al.* (2004), Hardie and McEnery (2003), Berglund (2000); links to these papers are available on this book's companion website.

Think about each study's approach to corpus linguistics. Where does it stand, in terms of the different criteria we introduced in this chapter? Remember, you are considering:

- The mode of communication of the corpus that the study uses;
- Whether it is (so-called) 'corpus-based' or 'corpus-driven' in its approach;
- Whether it uses a monitor corpus, a sample corpus or an opportunistic corpus;
- Whether it uses corpus annotations or not;
- Whether it complies with the principle of total accountability or not;
- Whether the corpus data is monolingual or multilingual.

(Q1-1) Imagine a situation where a study has been published that is generally agreed to mark a major advance in corpus linguistics. However, three years later, another study attempts to replicate the analysis and fails – in fact, it gets contradictory results. But the attempted replication was based on a different corpus with a different sampling frame, and a different set of computer programs was used to do the analysis. Obviously, these factors may have had an effect on the results.

How serious a problem would this situation be for the claims of the original study? For example, should researchers avoid any work that relies on its results, pending further replication studies? How often do we need to replicate a contested result before we can accept it as correct? How should we decide to apportion our efforts between replicating existing results versus establishing new results?