2

What is a corpus and what is in it?

2.1. CORPORA VS. MACHINE-READABLE TEXTS

Empirical research may be carried out using any written or spoken text. Indeed, such individual texts form the basis of many kinds of literary and linguistic analysis, for example, the stylistic analysis of a poem or novel or a conversation analysis of a television talk show. But the notion of a **corpus** as the basis for a form of empirical linguistics differs in several fundamental ways from the examination of particular texts. In principle, any collection of more than one text can be called a corpus: the term 'corpus' is simply the Latin for 'body', hence a corpus may be defined as any body of text. It need imply nothing more. But the term 'corpus' when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition provides for. These may be considered under four main headings:

sampling and representativeness

finite size

machine-readable form

a standard reference.

2.1.1. Sampling and representativeness

In linguistics, we are often more interested in a whole variety of a language, rather than in an individual text or author. In such cases we have two options for our data collection: first, we could analyse every single utterance in that variety; or second, we could construct a smaller sample of the variety. The first option is impracticable except in a very few cases, for example, with a dead language which has few extant texts. More often, the total text population is huge, and with a living language such as English or German the number of utterances is constantly increasing and theoretically infinite. To analyse every utterance in such a language would be an unending and impossible task. It is therefore necessary to choose the second option and build a sample of the language variety in which we are interested.

As we discussed in Chapter 1, it was Chomksy's criticism of early corpora that they would always be skewed: in other words, some utterances would be excluded because they are rare, other much more common utterances might be excluded simply by chance, and chance might also act so that some rare utterances were actually included in the corpus. Although modern computer technology means that nowadays much larger corpora can be collected than those Chomsky was thinking about when he made these criticisms, his criticism about the potential skewedness of a corpus is an important and valid one which must be taken seriously. However, this need not mean abandoning the corpus analysis enterprise. Rather, consideration of Chomsky's criticism should be directed towards the establishment of ways in which a much less biased and more generally repesentative corpus may be constructed.

In building a corpus of a language variety, we are interested in a sample which is maximally representative of the variety under examination, that is, which provides us with as accurate a picture as possible of the tendencies of that variety, including their proportions. We would not, for example, want to use only the novels of Charles Dickens or Charlotte Brontë as a basis for analysing the written English language of the mid-nineteenth century. We would not even want to base our sample purely on text selected from the genre of the novel. What we would be looking for are samples of a broad range of different authors and genres which, when taken together, may be considered to 'average out' and provide a reasonably accurate picture of the entire language population in which we are interested. We shall return in more detail to this issue of corpus representativeness and sampling in Chapter 3.

2.1.2. Finite size

As well as sampling, the term 'corpus' also tends to imply a body of text of a finite size, for example 1,000,000 words. This is not, however, universally so. At Birmingham University, for example, John Sinclair's COBUILD team have been engaged in the construction and analysis of a collection of texts known as a **monitor corpus**. A monitor corpus, which Sinclair's team often prefer to call simply a 'collection of texts' rather than a 'corpus', is an open-ended entity. Texts are constantly being added to it, so that it gets bigger and bigger as more samples are added. Monitor corpora are primarily of importance in lexicographic work, which is the main interest of the COBUILD group. They enable lexicographers to trawl a stream of new texts looking for the occurrence of new words or for changing meanings of old words. Their main advantages are: (1) the age of the texts, which is not static and means that very new texts can be included, unlike the synchronic 'snapshot' provided by finite corpora; and (2) their scope, in that a larger and much broader sample of the language can be covered. Their main disadvantage is that, because they are constantly changing in size and are less rigorously sampled than finite corpora, they are not such a reliable source of quantitative (as opposed to qualitative) data about a language. With the exception of the monitor corpus observed, though, it should be noted that it is more often the case that a corpus has a finite number of words contained in it. At the beginning of a corpus-building project, the research plan will set out in detail how the language variety is to be sampled, and how many samples of how many words are to be collected so that a predefined grand total is arrived at. With the Lancaster-Oslo/Bergen (LOB) corpus and the Brown corpus the grand total was 1,000,000 running words of text; with the British National Corpus (BNC) it was 100,000,000 running words. Unlike the monitor corpus, therefore, when such a corpus reaches the grand total of words, collection stops and the corpus is not thereafter increased in size. (One exception to this is the London-Lund corpus, which was augmented in the mid-1970s by Sidney Greenbaum to cover a wider variety of genres.)

2.1.3. Machine-readable form

It should also be noted that nowadays the term 'corpus' almost always implies the additional feature 'machine-readable'. For many years, the term 'corpus' could be used only in reference to printed text. But now things have changed, so that this is perhaps the exception rather than the rule. One example of a corpus which *is* available in printed form is *A Corpus of English Conversation* (Svartvik and Quirk 1980). This corpus represents the 'original' London-Lund corpus (i.e. minus the additional examples of more formal speech added by Sidney Greenbaum in the 1970s). Although these texts are also available in machine-readable form within the London-Lund corpus, this work is notable as it is one of the very few corpora available in book format. The appearance of corpora in book form is likely to remain very rare, though the Spoken English Corpus has recently appeared in this format (Knowles, Williams and Taylor 1996).

There is also a limited amount of other corpus data (excluding context-free frequency lists and so on, prepared *from* corpora) which is available in other media. A complete key-word-in-context concordance of the LOB corpus is available on microfiche and, with spoken corpora, copies of the actual recordings are sometimes available for, amongst other things, instrumental phonetic analysis: this is the case with the Lancaster/IBM Spoken English Corpus, but not with the London-Lund corpus.

Corpora which are machine-readable possess several advantages over the original written or spoken format. The first and most important advantage of machine-readable corpora, as noted in Chapter 1, is that they may be searched and manipulated in ways which are simply not possible with the other formats. For instance, a corpus in book format, unless pre-indexed, would need to be read cover to cover in order to extract all instances of the word *boot*: with a machine-readable corpus, this task may be accomplished in at most a few minutes using concordancing software, or even, slightly more slowly, simply using the search facility in a word processor. The second advantage of machine-readable corpora is that they can be swiftly and easily enriched with additional information. We shall turn to this issue of **annotation** later in this chapter.

2.1.4. A standard reference

Although it is not an essential part of the definition of a corpus, there is also often a tacit understanding that a corpus constitutes a standard reference for the language variety which it represents. This presupposes its wide availability to other researchers, which is indeed the case with many corpora such as the Brown corpus of written American English, the LOB corpus of written British English and the London-Lund corpus of spoken British English. The advantage of a widely available corpus is that it provides a yardstick by which successive studies may be measured. New results on related topics may, for example, be directly compared with published results (so long as the methodology is made clear) without the need for re-computation. A standard corpus also means that a continuous base of data is being used and thus variation between studies may be less likely to be attributed to differences in the data being used, and more to the adequacy of the assumptions and methodologies contained in the study.

So a corpus in modern linguistics, in contrast to being simply any body of text, might more accurately be described as a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration. However, the reader should be aware of the possibilities for deviation in certain instances from this 'prototypical' definition.

2.2. TEXT ENCODING AND ANNOTATION

Corpora may exist in two forms: **unannotated** (i.e. in their existing raw states of plain text) or **annotated** (i.e. enhanced with various types of linguistic information). Unannotated corpora have been, and are, of considerable use in language study, but the utility of the corpus is considerably increased by the provision of annotation. The important point to grasp about an annotated corpus is that it is no longer simply a body of text in which the linguistic information is implicitly present. For example, the part-of-speech information 'third person singular present tense verb' is always present implicitly in the form *loves*, but it is only retrieved in normal reading by recourse to our pre-existing knowledge of the grammar of English. By contrast, a corpus, when annotated, may be considered to be a repository of linguistic information, because the information which was implicit in the plain text has been made explicit through concrete annotation. Thus our example of *loves* might in an annotated corpus read 'loves_VVZ', with the code VVZ indicating that it is a third person singular present tense (Z) form of a lexical verb (VV). Such annotation makes it quicker and easier to retrieve and analyse information about the language contained in the corpus. We shall discuss part-of-speech and other forms of linguistic annotation further in section 2.2.2.3.