Unit 1 Corpus linguistics: the basics

1.1 Introduction

This unit sets the scene by addressing some of the basics of corpus-based language studies. We will first briefly review the history of corpus linguistics (unit 1.2). Then the term *corpus*, as used in modern linguistics, will be defined (unit 1.3). Following this is an explanation of why corpus linguists use computers to manipulate and exploit language data (unit 1.4). We will then compare the intuition-based approach and the corpus-based approach to language (unit 1.5), which is followed by an explanation of why corpus linguistics is basically a methodology rather than an independent branch of linguistics (unit 1.6). Finally, we will consider the corpus-based vs. corpus-driven debate (unit 1.7).

1.2 Corpus linguistics: past and present

Although the term *corpus linguistics* first appeared only in the early 1980s (cf. Leech 1992: 105), corpus-based language study has a substantial history. The corpus methodology dates back to the pre-Chomskyan period when it was used by field linguists such as Boas (1940) and linguists of the structuralist tradition, including Sapir, Newman, Bloomfield and Pike (cf. Biber and Finegan 1991: 207). Although linguists at that time would have used shoeboxes filled with paper slips rather than computers as a means of data storage, and the ‘corpora’ they used might have been simple collections of written or transcribed texts and thus not *representative* (see unit 2), their methodology was essentially ‘corpus-based’ in the sense that it was empirical and based on observed data. As McEnery and Wilson (2001: 2-4) note, the basic corpus methodology was widespread in linguistics in the early 20th century.

In the late 1950s, however, the corpus methodology was so severely criticized that it became marginalized, if not totally abandoned, in large part because of the alleged ‘skewedness’ of corpora (cf. Chomsky 1962; see McEnery and Wilson 2001: 5-13 for a more detailed discussion). Chomsky’s criticism was undoubtedly true when it was made. At that time, the size of ‘shoebox corpora’ was generally very small, and those corpora were used primarily for the study of distinguishing features in phonetics (Ling 1999: 240), though a few linguists of this era, notably Jesperson (1909-1949) and Fries (1952), also used paper-based corpora to study grammar. Using paper slips and human hands and eyes, it was virtually impossible to collate and analyze large bodies of language data. Consequently the corpora of the time could rarely avoid being ‘skewed’.

But with developments in technology, and especially the development of ever more powerful computers offering ever increasing processing power and massive storage at relatively low cost, the exploitation of massive corpora became feasible. The marriage of corpora with computer technology rekindled interest in the corpus methodology. The first *modern corpus* (cf. unit 1.3) of the English language, the Brown corpus (i.e. the Brown University Standard Corpus of Present-day American English, see unit 7.4), was built in the early 1960s for American English. Since then, and increasingly so from the 1980s onwards, the number and size of corpora and corpus-based studies have increased dramatically (cf. Johansson 1991: 12). Nowadays, the corpus methodology enjoys widespread popularity. It has opened up or foregrounded many new areas of research. Much of the research presented in this book would not have been produced without corpora. Unsurprisingly, as we will see in unit 10, corpora have revolutionized nearly all branches of linguistics.

1.3 What is a corpus?

In modern linguistics, a corpus can be defined as a body of naturally occurring language, though strictly speaking:

> It should be added that computer corpora are rarely haphazard collections of textual material: They are generally assembled with particular purposes in mind, and are often assembled to be (informally speaking) *representative* of some language or text type. (Leech 1992: 116)

Sinclair (1996) echoes Leech's definition of *corpus*, as he also stresses the importance of *representativeness* (see unit 2): 'A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.' The 'linguistic criteria', which are external to the texts themselves and dependent upon the intended use for the corpus (cf. Aston and Burnard 1998: 23; see units 2 and 11 for further discussion), are used to select and put together these texts 'in a principled way' (Johansson 1998: 3). Thus a corpus is different from a random collection of texts or an archive whose components are unlikely to have been assembled with such goals in mind (Aston and Burnard 1998:5; Leech and Fligelstone 1992: 120). Rather, the term *corpus* as used in modern linguistics can best be defined as a collection of sampled texts, written or spoken, in machine readable form which may be annotated with various forms of linguistic information (see unit 4 for a discussion of corpus annotation).

There are many ways to define a corpus (e.g. Francis 1992: 17; Atkins, Clear and Ostler 1992: 1), but there is an increasing consensus that a corpus is a collection of (1) *machine-readable* (2) *authentic* texts (including transcripts of spoken data) which is (3) *sampled* to be (4) *representative* of a particular language or language variety. While all scholars agree upon the first two qualities, there are differing opinions regarding what can be counted as representative. Also, the question of what sampling techniques should be used to achieve representativeness is contentious. While some scholars propose that a corpus must be defined in linguistic terms (e.g. the distribution of words or other patterns), it is our view that non-linguistic (or extralinguistic) parameters should be used as important definitional criteria also (see units 2 and 11 for further discussion).

It has been argued that corpora like the Lancaster Corpus of Abuse (i.e. LCA, see McEnery, Baker and Hardie 2000: 46), which are built using extracts from large corpora to study a specific linguistic phenomenon, are not corpora in a real sense. Such an argument, nevertheless, is arguably misleading for a number of reasons. First, corpora of this kind certainly meet the four criteria of a corpus as discussed above. Second, not all corpora are *balanced*. Specialized corpora serve a different yet important purpose from balanced corpora (see unit 7.3). If specialized corpora which are built using a different sampling technique from those for balanced corpora were discounted as 'non-corpora', then corpus linguistics would have contributed considerably less to language studies. Third, it is simply unreasonable to argue that a subcorpus, which contains part of a larger corpus, is not a corpus. In fact, some corpus tools helpfully allow users to define a subcorpus from a larger corpus. For example, SARA (Aston and Burnard 1998) allows users to define subcorpora from the British National Corpus (i.e. BNC, see unit 7.2) using the selected parameters; Xaira, the XML-aware version of SARA (cf. Burnard and Todd 2003; see unit 3.3 for a discussion of Extensible Markup Language), even allows users to define a subcorpus from a large corpus through a query. The new version of WordSmith Tools (version 4, see Scott 2003) now includes a WebGetter function to help users to build their corpora using web pages on the Internet which contain the specified search patterns. If carefully selected subcorpora do not merit the label 'corpus', then corpora (or subcorpora) built using the corpus tools outlined above would not be called corpora either. So while it may be appealing to define precisely what a corpus is, the criteria should not be applied with such zeal that terminology is used as a tool to exclude carefully composed collections of language data from corpus-based research. The term *corpus*, while useful, should always be viewed as a somewhat vague and inclusive term.

1.4 Why use computers to study language?

It is clear from the previous section that the essential qualities of a corpus include machine-readability, authenticity and representativeness. Authenticity will be discussed when we compare a corpus-based and intuition-based approach whilst representativeness, together with related issues such as balance and sampling, will be explored in units 2 and 11. In this section, we will focus on machine-readability and explain why corpus linguists use computers to manipulate and exploit language data.

Machine-readability is a *de facto* attribute of modern corpora. Electronic corpora have advantages unavailable to their paper-based equivalents. The most obvious advantage of using a computer for language study is the speed of processing it affords and the ease with which it can manipulate data (e.g. searching, selecting, sorting and formatting). Computerized corpora can be processed and manipulated rapidly at minimal cost. Secondly, computers can process machine-readable data accurately and consistently (cf. Barnbrook 1996: 11; see also unit 4.2). Thirdly, computers can avoid human bias in an analysis, thus making the result more reliable. Finally, machine-readability allows further automatic processing to be performed on the corpus so that corpus texts can be enriched with various metadata and linguistic analyses (see units 3 and 4). It is the use of computerized corpora, together with computer programs which facilitate linguistic analysis, that distinguishes modern machine-readable corpora from early 'drawer-cum-slip' corpora (Svartvik 1992: 9). Without computers, many of the corpus-based studies undertaken in the past 20 years would have been impossible. As Tognini-Bonelli (2000: 210) observes, the computer has affected the methodological frame of linguistic enquiry. Given the prominence of the computer in corpus linguistics, it is unsurprising that corpora are typically in fact computerized corpora, and 'computer corpus linguistics' (CCL) has been suggested as the improved name for corpus linguistics (CL) (Leech 1992: 106). However, CCL is not a term that is widely used, as most scholars assume that CL implies CCL.

1.5 The corpus-based approach vs. the intuition-based approach

In principle, by using the intuition-based approach, researchers can invent purer examples instantly for analysis, because intuition is readily available and invented examples are free from language-external influences existing in naturally occurring language. However, intuition should be applied with caution (cf. Seuren 1998: 260-262). Firstly, it is possible to be influenced by one's dialect or sociolect; what appears unacceptable to one speaker may be perfectly felicitous to another. Assuming that what we see in a corpus is largely grammatical and/or acceptable, the corpus at least provides evidence of what speakers believe to be acceptable utterances in their language, typically free of the overt judgment of others. Secondly, when one invents an example to support or disprove an argument, one is consciously monitoring one's language production. Therefore, even if one's intuition is correct, the utterance may not represent typical language use. The corpus-based approach, in contrast, draws upon authentic or real texts, though authenticity itself may be a cause of dispute (see units 10.8 and 12). Finally, results based on introspection alone are difficult to verify as introspection is not observable. All of these disadvantages are circumvented by the corpus-based approach. Additional advantages of the corpus-based approach are that a corpus can find differences that intuition alone cannot perceive (cf. Francis, Hunston and Manning 1996; Chief, Hung, Chen, Tsai and Chang 2000), and a corpus can yield reliable quantitative data.

Broadly speaking, compared with the more traditional intuition-based approach, which rejected or largely ignored corpus data, the corpus-based approach can offer the linguist improved reliability because it does not go to the extreme of rejecting intuition while attaching importance to empirical data. The key to using corpus data is to find the balance between the use of corpus data and the use of one's intuition. As Leech (1991: 14) comments:

> Neither the corpus linguist of the 1950s, who rejected intuition, nor the general linguist of the 1960s, who rejected corpus data, was able to achieve the interaction of data coverage and the insight that characterise the many successful corpus analyses of recent years.

While the corpus-based approach has obvious advantages over a purely intuition-based approach, not all linguists accept the use of corpora, as we will see in unit 12. Indeed, it must be accepted that not all research questions can be addressed by the corpus-based approach (cf. unit 10.15). This in large part explains why the corpus-based approach and the intuition-based approach are not mutually exclusive. The two are complementary and must be so if as broad a range of research questions as possible are to be addressed by linguists (cf. McEnery and Wilson 2001: 19; Sinclair 2003: 8).

1.6 Corpus linguistics: a methodology or a theory?

We have, so far, assumed that corpus linguistics is a methodology rather than an independent branch of linguistics. This view, however, is not shared by all scholars. For example, it has been argued that corpus linguistics 'goes well beyond this methodological role' and has become an independent 'discipline' (Tognini-Bonelli 2001: 1). While we agree that corpus linguistics is 'really a domain of research' and 'has become a new research enterprise and a new philosophical approach to linguistic enquiry' (*ibid*), we maintain that corpus linguistics is indeed a methodology rather than an independent branch of linguistics in the same sense as phonetics, syntax, semantics or pragmatics. These latter areas of linguistics describe, or explain, a certain aspect of language use. Corpus linguistics, in contrast, is not restricted to a particular aspect of language. Rather, it can be employed to explore almost any area of linguistic research (see unit 10). Hence, syntax can be studied using a corpus-based or non-corpus-based approach; similarly, we have corpus semantics and non-corpus semantics.

As corpus linguistics is a whole system of methods and principles of how to apply corpora in language studies and teaching/learning, it certainly has a theoretical status. Yet theoretical status is not theory itself. The qualitative methodology used in social sciences also has a theoretical basis and a set of rules relating to, for example, how to conduct an interview, or how to design a questionnaire, yet it is still labelled as a methodology upon which theories may be built. The same is true of corpus linguistics.

Definitional confusion bedevils corpus linguistics. As we have seen with the term *corpus* itself, strict definitions often fail to hold when specific examples are considered. Similarly, with the methodology question, the attempt to construct corpus linguistics as anything other than a methodology ultimately fails. In fact, even those who have strongly argued that corpus linguistics is an independent branch of linguistics have frequently used the terms ‘approach’ and ‘methodology’ to describe corpus linguistics (e.g. Tognini-Bonelli 2001). Hence, as with the term *corpus* itself, our approach is to take the less rigid, indeed less limiting, position. Corpus linguistics should be considered as a methodology with a wide range of applications across many areas and theories of linguistics.

1.7 Corpus-based vs. corpus-driven approaches

One further, notable, area where differences emerge between corpus linguists is with regard to the question of corpus-based and corpus-driven approaches. In the corpus-based approach, it is said that corpora are used mainly to ‘expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study’ (Tognini-Bonelli 2001: 65). Corpus-based linguists are accused of not being fully and strictly committed to corpus data as a whole as they have been said to discard inconvenient evidence (i.e. data not fitting the pre-corpus theory) by ‘insulation’, ‘standardisation’ and ‘instantiation’, typically by means of annotating a corpus (see unit 4). In contrast, corpus-driven linguists are said to be strictly committed to ‘the integrity of the data as a whole’ (*ibid*: 84) and therefore, in this latter approach, it is claimed that ‘[t]he theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus’ (*ibid*: 85). However, the distinction between the corpus-based vs. corpus-driven approaches is overstated. In particular the latter approach is best viewed as an idealized extreme. There are four basic differences between the corpus-based vs. corpus-driven approaches: types of corpora used, attitudes towards existing theories and intuitions, and focuses of research. Let us discuss each in turn.

Regarding the type of corpus data used, there are three issues – representativeness, corpus size and annotation. Let us consider these in turn. According to corpus-driven linguist, there is no need to make any serious effort to achieve corpus balance and representativeness (see unit 2) because the corpus is said to balance itself when it grows to be big enough as the corpus achieves so-called cumulative representativeness. This initial assumption of self-balancing via cumulative representativeness, nonetheless, is arguably unwarranted (cf. unit 2.4). For example, one such cumulatively representative corpus is a corpus of Zimbabwean English Louw (1991) used in his contrastive study of collocations of in British English and Zimbabwean English. This study shows that the collocates of *wash* and *washing*, etc in British English are *machine, powder* and *spin* whereas in Zimbabwean English the more likely collocates are *women*, *river*, *earth* and *stone*. The different collocational behaviours were attributed to the fact that the Zimbabwean corpus has a prominent element of literary texts such as Charles Mungoshi’s novel *Waiting for the Rain*, ‘where women washing in the river are a recurrent theme across the novel’ (Tognini-Bonelli 2001: 88). One could therefore reasonably argue that this so-called cumulatively balanced corpus was skewed. Especially where whole texts are included, a practice corpus-driven linguists advocate, it is nearly unavoidable that a small number of texts may seriously affect, either by theme or in style, the balance of a corpus (see units 2.5 and 11.4 for a further discussion of whole texts). Findings on the basis of such cumulatively representative corpora may not be generalizable beyond the corpora themselves as their representativeness is highly idiosyncratic.

The corpus-driven approach also argues for very large corpora. While it is true that the corpora used by corpus-driven linguists are very large (for example, the latest release of the Bank of English has grown to 524 million words as of early 2004), size is not all-important (see unit 8.2), as Leech (1991: 8-29) notes (cf. also McCarthy and Carter 2001). Another problem for the corpus-driven approach relates to frequency. While it has been claimed that in the corpus-driven approach corpus evidence is exploited fully, in reality frequency may be used as a filter to allow the analyst to exclude some data from their analysis. For example, a researcher may set the minimum frequency of occurrence for a pattern which it must reach before it merits attention, e.g. it must occur at least twice – in separate documents (Tognini-Bonelli 2001: 89). Even with such a filter, a corpus-driven grammar would consist of thousands of patterns which would bewilder the learner. It is presumably to avoid such bewilderment that the patterns reported in the *Grammar Patterns* series (Francis, Hunston and Manning 1996, 1998), which are considered as the first results of the corpus-driven approach, are not even that exhaustive. Indeed, faced with the great number of concordances, corpus-driven linguists often analyze only the n^{th} occurrence from a total of X instances. This is in reality currently the most practical way of exploring a very large unannotated corpus. Yet if a large corpus is reduced to a small dataset in this way, there is little advantage in using very large corpora to explore frequent features. It is also difficult to see how it can be claimed that the corpus data is exploited fully and the integrity of the data is respected in such cases. It appears, then, that the corpus-driven approach is not so different from the corpus-based approach – while the latter allegedly insulates theory from data or standardizes data to fit theory, the former filters the data via apparently scientific random sampling, though there is no guarantee that the corpus is not explored selectively to avoid inconvenient evidence.

The corpus-driven linguists have strong objections to corpus annotation. This is closely associated with the second difference between the two approaches – attitudes towards existing theories and intuitions. It is claimed that the corpus-driven linguists come to a corpus with no preconceived theory, with the aim of postulating linguistic categories entirely on the basis of corpus data, though corpus-driven linguists do concede that pre-corpus theories are insights cumulated over centuries which should not be discarded readily and that intuitions are essential in analyzing data. This claim is a little surprising, as traditional categories such as nouns, verbs, prepositions, subjects, objects, clauses, and passives are not uncommon in studies which identify themselves as corpus-driven. When these terms occur they are used without a definition and are accepted as given. Also, linguistic intuitions typically come as a result of accumulated education in preconceived theory. So applying intuitions when classifying concordances may simply be an implicit annotation process, which unconsciously makes use of preconceived theory. As implicit annotation is not open to scrutiny, it is to all intents and purposes unrecoverable and thus more unreliable than explicit annotation (see unit 4.2). Corpus-based linguists do not have such a hostile attitude toward existing theory. The corpus-based approach typically has existing theory as a starting point and corrects and revises such theory in the light of corpus evidence. As part of this process, corpus annotation is common. Annotating a corpus, most notably through part-of-speech tagging (see unit 4.4.1), inevitably involves developing a set of parts of speech on the basis of an existing theory, which is then tested and revised constantly to mirror the attested language use. In spite of the clear usefulness of outcomes of corpus annotation, which greatly facilitate corpus exploration, the process of annotation itself is also important. As Aarts (2002: 122) observes, as part of the annotation process the task of the linguist becomes 'to examine where the annotation fits the data and where it does not, and to make changes in the description and annotation scheme where it does not.' The claimed independence of preconception on the part of corpus-driven linguists is clearly an overstatement. A truly corpus-driven approach, if defined in this way, would require something such as someone who has never received any education related to language use and therefore is free from preconceived theory, for as Sampson (2001: 135) observes, schooling also plays an important role in forming one's intuitions. Given that preconceived theory is difficult to totally reject and dismiss, and intuitions are indeed called upon in corpus-driven linguistics, it is safe to conclude that there is no real difference between the corpus-driven demand to re-examine pre-corpus theories in the new framework and corpus-based linguists' practice of testing and revising such theories. Furthermore, if the so-called proven corpus-driven categories in corpus-driven linguistics, which are supposed to be already fully consistent with and directly reflect corpus evidence, also need refinement in the light of different corpus data, the original corpus data is arguably not representative enough. The endless refinement will result in inconsistent language descriptions which will place an unwelcome burden on the linguist or the learner. In this sense, the corpus-driven approach is no better than the corpus-based approach.

The third important difference between the corpus-driven and corpus-based approaches is their different research focuses. As the corpus-driven approach makes no distinction between lexis, syntax, pragmatics, semantics and discourse (because all of these are pre-corpus concepts and they combine to create meaning), the holistic approach provides, unsurprisingly, only one level of language description, namely, the functionally complete unit of meaning or language patterning. In studying patterning, corpus-driven linguists concede that while collocation can be easily identified in KWIC concordances of unannotated data, colligation is less obvious unless a corpus is grammatically tagged. Yet a tagged corpus is the last thing the corpus-driven linguists should turn to, as grammatical tagging is based on preconceived theory, and consequently results in a loss of information, in their view. To overcome this problem, Firth's definition of colligation is often applied in a loose sense – in spite of the claim that corpus-driven linguists is deeply rooted in Firth's work – because studying colligation in Firth's original sense necessitates a tagged or even a parsed corpus. According to Firth (1968: 181), colligation refers to the relations between words at the grammatical level, i.e. the relations of 'word and sentence classes or of similar categories' instead of 'between words as such.' But nowadays the term colligation has been used to refer not only to significant co-occurrence of a word with grammatical classes or categories (e.g. Hoey 1997, 2000; Stubbs 2001c: 112) but also to significant co-occurrence of a word with grammatical words (e.g. Krishnamurthy 2000). The patterning with grammatical words, of course, can be observed and computed even using a raw corpus.

A final contrast one can note between the corpus-based and corpus-driven approaches is that the corpus-based approach is not as radical as the corpus-driven approach. The corpus-driven approach claims to be a new paradigm within which a whole language can be described. No such claim is entailed in the corpus-based approach. Yet as we will see in unit 10, the corpus-based approach, as a methodology, has been applied in nearly all branches of linguistics.

The above discussion shows that the sharp distinction between the corpus-based vs. corpus-driven approaches to language studies is in reality fuzzy. As with the definition of what a corpus is and the theory vs. methodology distinction, we maintain a less rigid distinction between the two approaches. In our book, the term corpus-based is used in a broad sense, encompassing both corpus-based and corpus-driven approaches, as suggested by the title of this book.

1.8 Unit summary and looking ahead

This unit addressed some basic issues in corpus linguistics, including a brief review of the history of corpus linguistics, a definition of *corpus* as used in modern linguistics, a discussion of the advantages of using computers in language studies, a comparison of the intuition-based and the corpus-based approaches, an explanation of why corpus linguistics should be viewed as a methodology rather than an independent branch of linguistics, and finally a discussion of the debate over the corpus-based vs. corpus-driven linguistics.

In this unit, we focused only on one salient feature of a modern corpus, namely, machine-readability. Other issues of corpus design (e.g. balance, representativeness, sampling and corpus size) will be discussed in units 2 and 8, and further explored in unit 11. Corpus processing (e.g. data capture, corpus markup and annotation) will be discussed in units 3 – 4 and 8. Using corpora in language studies will be introduced in unit 10 and further discussed in Section B and explored in Section C of this book.