2 *The corpus as object: Design and purpose*

As corpora have become larger and more diverse, and as they are more frequently used to make definitive statements about language, issues of how they are designed have become more important. Four aspects of corpus design are discussed in this chapter: size, content, representativeness and permanence. The chapter also summarises some types of corpus investigation, each of which treats the corpus as a different kind of object.

Issues in corpus design

Size

As computer technology has advanced since the 1960s, so it has become feasible to store and access corpora of ever-increasing size. Whereas the LOB Corpus and Brown Corpus seemed as big as anyone would ever want, at the time, nowadays 1 million words is fairly small in terms of corpora. The British National Corpus is 100 million words; the Bank of English is currently about 400 million. CANCODE is 5 million words. The feasible size of a corpus is not limited so much by the capacity of a computer to store it, as by the speed and efficiency of the access software. If, for example, counting the number of past and present tense forms of the verb *BE* in a given corpus takes longer than a few minutes, the researcher may prefer to use a smaller corpus whose results might be considered to be just as reliable but on which the software would work much more speedily.

Sheer quantity of information can be overwhelming for the observer. In the Bank of English corpus, for example, a search for the word *point* gives almost 143,000 hits. Few researchers can obtain useful information simply by looking at so many concordance lines. One solution is to use a smaller corpus, which will give less data. Another is to keep the large corpus, but to use software which will make a selection of data from the whole, either by selecting at random a proportion of the total concordance lines, or by identifying and allowing selection of the most significant collocates, or other

significant features. In this way, the data from a very large corpus can be reduced to a manageable scale whilst retaining the advantages of coverage of the large corpus.

The question of corpus size can be a contentious one. There are many arguments to support the view that a small corpus *can* be valuable under certain circumstances. Carter and McCarthy (1995: 143), for example, argue that for the purposes of studying grammar in spoken language a relatively small corpus is sufficient (because grammar words tend to be very frequent). Some kinds of corpus investigation that depend on the manual annotation of the corpus are also, necessarily, restricted in terms of the size of the corpus that can feasibly be annotated. Similarly, many commercially available concordance software packages restrict the size of the corpus they can be used with (see Biber et al 1998: 285–286 for details). Someone aiming to build a balanced corpus may restrict the amount of data from one source in order to match data from another source (for example, the amount of written data may be kept smaller than it need be because it is difficult to obtain large amounts of spoken data).

A more extreme view is that for some purposes there is an *optimum* corpus size that should not be exceeded even if practical considerations would allow it, that is, that a relatively small corpus is not just sufficient but also necessary. However, this argument usually refers to the difficulties of processing large amounts of information about very frequent words, as discussed above (e.g. Carter and McCarthy 1995: 143). The opposing view would be that it is preferable to select from a large amount of data than to restrict the amount of data available (Sinclair 1992).

Arguments about optimum corpus size tend to be academic for most people. Most corpus users simply make use of as much data as is available, without worrying too much about what is not available. As well as the very large, general corpora designed to assist in writing dictionaries and other reference books, there are thousands of smaller corpora around the world, some comprising only a few thousand words and designed for a particular piece of research.

Content

It is a truism that a corpus is neither good nor bad in itself, but suited or not suited to a particular purpose. Decisions about what should go into a corpus are based on what the corpus is going to be used for, but also about what is available. A researcher who wishes to study

academic articles in History, for example, may plan to build a corpus consisting simply of published articles. Even with such a straightforward design plan, however, a number of issues will have to be resolved, such as:

- Will the articles come from one journal, or from a range of journals? How will the journal(s) be selected?
- Will the articles chosen be restricted to those apparently written by native speakers of the target language, or not?
- Will articles other than research articles, such as book reviews, be included?
- If the articles are to be stored electronically, and copyright clearance is required, which publishers are willing to give permission for their articles to be used?
- If the articles are required in electronic form, which publishers are willing to make electronic versions of their journals available?

To the first three of these questions, there are no 'right' answers. The best choice in each case will depend on the precise purpose of the research. The last two questions refer to purely pragmatic issues, but these may in the end affect the design of the corpus more profoundly than the first three.

Publishers' policies are not the only issues affecting availability for corpora. Parallel corpora, for example, are composed of texts that are translations of each other. They can make use only of texts that have been translated between the two languages involved. Other issues of selection come second to the question of what translations exist. Using unpublished material in a corpus does not obviate the issue of availability. For example, a researcher wishing to study student writing may ask students to submit electronic versions of their essays to assist in this work. If this submission is voluntary, the researcher will collect only the essays whose authors are willing for their essays to be used.

For some purposes, corpus design is scarcely an issue. For example, if a corpus is being used to encourage learners to investigate language data for themselves, the precise contents of that corpus may be relatively unimportant. Any collection of newspapers written in the target form of the language, for example, will be adequate for some investigation purposes, even if the newspapers do not represent all aspects of language use (see, for example, Dodd 1997). Where the corpus is being taken as representative of a language or a language variety, the notion of design and balance is very important. This will be considered separately below.

Balance and representativeness

Corpora are very often intended to be representative of a particular kind of language. If the object of study is academic prose, or casual conversation, or the language of newspapers, or American English, for example, an attempt must be made to build a corpus that is representative of the whole. (See Kennedy 1991: 52; Crowdy 1993; Burnard ed. 1995 for discussions of representativeness in the British National Corpus.) Usually, this involves breaking the whole down into component parts and aiming to include equal amounts of data from each of the parts. For example, under 'the language of newspapers', it may be decided to include a range of newspaper types (broadsheet and tabloid, for instance) and a range of article types (hard news, human interest, editorials, letters, sports, business, advertisements and so on). A balanced corpus might be said to consist of equal numbers of words in each category: broadsheet hard news; tabloid hard news; broadsheet human interest; tabloid human interest and so on. This notion of balance is, however, open to question. It might be argued that broadsheet newspapers contain more words, on average, than tabloids, and that therefore the corpus should contain more from the broadsheets than from the tabloids. Alternatively, it might be argued that tabloids are read by more people than broadsheets and that therefore they should comprise a larger part of the corpus. If the broadsheets contain more hard news and editorial than the tabloids, and the tabloids contain more human interest and sports news, should the proportions in the corpus be adjusted accordingly? In the case of newspapers, the solution may be to include all issues of a selection of publications from a given week, month or year. This will allow the proportions to determine themselves. It does mean, however, that more data will probably be gathered from one type of publication than another. Similar problems with relation to the collection of academic prose or casual conversation, for example, are not open to so easy a solution. The problem is that 'being representative' inevitably involves knowing what the character of the 'whole' is. Where the proportions of that character are unknowable, attempts to be representative tend to rest on little more than guesswork.

The problem of representation and balance becomes more difficult when the corpus is supposed to represent a regional variety of English (Singapore English, or British English, or Australian English), with all its complexity of internal variation. In practice, the way that this is done depends on the purpose of the corpus. The International Corpus of English (ICE), for example, exists to facilitate comparison

between regional varieties. To enable this to be done, the compilers of each regional corpus were given precise instructions as to how much of what kinds of language (newspapers, literature, conversation etc) to include, irrespective of the internal make-up of their own variety. The Bank of English is designed as a resource for writers of dictionaries and other reference books for learners of English. It is assumed that learners will expect to find standard English described in such books, and as a result there is no attempt to represent regional or social varieties of spoken English in it. There is in fact a good deal of non-standard spoken language in the BoE, but it is not possible to use this corpus to compare, say, the English spoken in Bristol with the English spoken in Newcastle. The British National Corpus, on the other hand, is designed to represent the different varieties of spoken English within Britain, and to allow comparisons to be made, although only between very broadly identified groups, such as 'south' and 'north' (Rayson et al 1997).

Spoken language exists in unknowable quantities and in an unknowable range of varieties. How, then, is a corpus builder to 'represent' the diversity in a meaningful way? One approach is to make a list of variables, as discussed above, taking into account age, gender, social class and home town of each speaker, as well as settings or genres, such as casual conversation, service encounter, radio broadcast, classroom, office and so on. Roughly equal amounts of data can then be collected under each heading. This will give a reasonable spread of registers, but it is important to remember that we do not know how balanced or representative the resulting corpus is. For example, we do not know how much weight should be given to 'service encounters' as opposed to 'office talk', and so on, or, indeed, how many and which genres should be included. After all, no definitive list of spoken genres exists. An alternative is to include whatever data is available, covering as many different settings and genres as possible, but without attempting to balance the corpus between the types. The advantage of this is that no data has to be wasted, as conversations between male speakers, for example, do not have to be discarded because no conversations of an equivalent kind between female speakers have been obtained. The hope is that once the corpus is of a substantial size the relevant figures can be checked and efforts made to collect data from under-represented groups, so that balance, where it is possible, is achieved after the corpus is (partially) complete, rather than from the outset. The disadvantage is that even where there is adequate information on the relevant groups in the population as a whole (for example, there are roughly equal numbers of women and men) the corpus may be biased towards one

group or the other. The books in the original Cobuild corpus, for example, consisted of 49 written by men and only 15 written by women (Renouf 1987: 33).

The real question as regards representativeness is how the balance of a corpus should be taken into account when interpreting data from that corpus. De Beaugrande (1999), for example, reports that the Bank of English (the 211 million word version) contained 11 instances of *did not/didn't mean to kill*, whereas the British National Corpus contained only 2 instances. A reviewer of de Beaugrande's paper (1999: 258) apparently commented that this discrepancy might arise because the Bank of English contains proportionally more journalistic prose than the British National Corpus does. It is not possible to reconstruct from de Beaugrande's note what the full import of the reviewer's comment was. It is most likely to be a warning: 'Don't assume that *didn't mean to kill* is typical in English as a whole, because it is generally only found in newspapers.' This is true: of the 23 instances of *did not/didn't mean to kill* in the current Bank of English, four occur in books and all the others in newspapers. On the other hand, this does not substantially affect de Beaugrande's point, which is that *didn't mean to* often has a pragmatic meaning of apology rather than being a straightforward statement of fact, *I didn't mean to kill him* being only one instance of this, if a dramatic one. Corpus design is less important here than, where necessary, paying attention to the distribution of a phrase across sub-corpora. This can be done by looking at comparative frequencies, even where sub-corpora are of unequal sizes. For example, the phrase *didn't mean to* (followed by any verb) is found most frequently in books (about 3 times per million words), then in spoken English (1.3 times per million words), then in journalism (below once per million words). The relative frequency of *did not mean to kill* in journalism is explained by the large number of reports of court cases in that register.

Permanence

One aspect of representativeness that is sometimes overlooked is the diachronic aspect. Any corpus that is not regularly updated rapidly becomes unrepresentative, in the sense that it no longer represents the language as currently written or spoken. A monitor corpus, which is added to very frequently, clearly has temporal representativeness as a key aspect of its design. It is usually, however, impossible to maintain a monitor corpus that also includes texts of many different types, as some are just too expensive or time-consuming to

collect on a regular basis. On the other hand, the easy availability of newspaper material makes it feasible to build a monitor corpus that can be enlarged and updated annually, weekly, or even daily (for an account of monitor corpora see Sinclair 1991: 23–26). Such a corpus does not represent all kinds of language use, but can be used to keep track of changes in the language appearing in newspapers. There is some evidence (Hundt and Mair 1999) that newspapers, although they have their own style, are a good source of general information about language change, as they incorporate changes more quickly than other kinds of discourse do. A monitor corpus comprising only newspaper data sacrifices 'what is desirable' (adding to a general corpus from all of its component registers every year) to 'what is feasible' (developing a corpus that is restricted in terms of register but expansive in size and in currency).[1]

A monitor corpus is an object in constant flux, something that is transient and fleeting. The other side of the coin is a corpus which is a permanent artefact, a definitive and fixed representation of a language variety. Permanence is significant when a corpus has symbolic value. For many minority languages, the establishment of a corpus serves to assert identity and importance, rather as writing a dictionary of the language has always done. For example, the TRACTOR archive (a resource produced by the Trans-European Language Resources Infrastructure research project) contains corpora of languages from newly independent European countries, such as Estonian, Latvian, Lithuanian, Slovenian, Ukrainian and Uzbek, as well as the more widely spoken languages, such as English, French and Spanish (http://www.telri.de). In the Netherlands, a corpus of Dutch from the eighth to the twenty-first century is sponsored by the Dutch and Belgian governments (Kruyt 2000). The motivation behind the financing of such a project is political as well as linguistic.

A corpus, once compiled, can assist in other language-maintenance projects such as the preparation of descriptions of the vocabulary, grammar and pragmatic usage of contemporary native speakers. Ahmed and Davies (1997: 158), who have compiled corpora of Welsh, refer to these products of corpus research as 'an infrastructure for promoting Welsh as a language'. They note that the preparation of 'term banks' (lists of technical terms from science and technology in the target language) is particularly important to the maintenance

[1] I am indebted to conversations with Professor Wolfgang Teubert for the development of this idea.

of a minority language such as Welsh, and give details of how this can be done using a corpus as a resource.[2]

The issue of permanence raises the question of what kind of object a corpus is. At extreme ends of the spectrum, it is a permanent, definitive record of a language, or a fluctuating trace of change. Further, different views about the nature of a corpus as an object will be considered in the next section.

Corpus, text and language

Another basic distinction that needs to be made is between a corpus as a collection of texts and a corpus as a collection of samples of language (Scott 2000). In terms of how the corpus is designed, this is a fairly simple distinction: a corpus may consist of whole texts or of extracts from texts. The distinction goes beyond this, however, to how a corpus is investigated, and for what purpose. Below are some examples of studies, each of which takes a different approach to what a corpus is.

The corpus as a collection of texts I

For a study of *thank you* as a conversational closing, Aston (1995) has collected two corpora consisting of 'naturally occurring service encounters between assistants and customers' in bookshops in Britain and in Italy (Aston 1995: 64). There are 160 English encounters and 181 Italian ones. Fifty-three per cent of the English encounters include an expression of thanks, while 70% of the Italian ones do. Of the English thanks, 94% are produced by the customer, while 85% of the Italian ones are (1995: 66). Through close examination of the encounters with thanks, Aston gives an account of how and why thanking is employed, arguing that its use goes far beyond a simple politeness token and that 'thanking can be seen as motivated to a large extent by concerns of conversational management, where there is a need to ratify referential and/or role alignment' (1995: 78). For example, he cites examples from the corpus where a customer needs

[2] An issue of current importance is the status of sign languages, such as British Sign Language, which is at the time of writing not recognised as a language by the British government. A corpus of sign language presents particular challenges, as the language does not have sounds that can be represented by letters. However, building a corpus of BSL, which would necessitate annotating a video corpus with transcription symbols (see, for example, Brien et al 1992: xvii–xxiii), would aid the cause of those pressing to have BSL officially recognised.

to check that he/she has understood the assistant's information correctly. Once the problem of possible misunderstanding has been settled, the customer typically uses thanking in acknowledgement (1995: 70).

The research that Aston reports in this article uses methods which essentially belong to conversation analysis. Although the corpus may be stored electronically, the only part of the method that could use corpus search techniques is the identification of which of the candidate dialogues include thanking and therefore form part of this study. In this particular paper, then (though not in his other work), Aston uses an electronic corpus simply as a convenient way to store a collection of texts. (For other work on thanking and other routines, see Aijmer 1996.)

More recently, studies have combined techniques of conversation analysis with the possibilities of annotation that corpora provide. For example, Barth (1999) discusses examples of the concessive relation in spoken English. She argues that there is a 'cardinal' concessive pattern consisting of (X) a statement, (X') a positive reference to that statement, (Y) a counter-argument; the pattern may occur in the canonical order or with variation. This is illustrated in the following two examples (heavily adapted and simplified from Barth 1999, with intonation coding omitted):

Code	Dialogue	Explanatory gloss
(1)		
(X)	S: and it was it was like the norm. Everybody knew that that was really a sacred thing to a certain extent	[in the 1940s, everyone knew that sexual abstinence was the norm]
(X')	D: right	[agree that everybody knew that]
(Y)	and nobody did it though	[people did not practise sexual abstinence]
(2)		
(X)	R: but when we look at . . . your inner circle of advisors we see white men only	[the government employs only (white) men]
(Y)	B: you can look all around and you'll see first class strong women	[the government employs women]
(X')	Uh Jim Baker's a man yeah I agree . . .	[one top advisor is a man]
(Y)	but look around who's . . . around with him there	[the others are women]

Like Aston, Barth uses the corpus as a convenient way of storing texts, with the additional advantage that selected features can be tagged. The annotation is, however, carried out entirely by hand.

According to Barth, the concession sequences are identified by the presence of connectors such as *but* and *though*, by paralinguistic cues such as intonation, and by contextual cues, that is 'world knowledge'. Unfortunately, neither connectors nor particular intonation patterns can reliably identify only these sequences in a corpus. Items such as *but* and *though* signal other forms of contrast, for example. The 'world knowledge' in the above examples seems to consist of the recognition of contrasts, such as that between *everybody knew that* and *nobody did it* and that between *men* and *women*, and the recognition of class membership, e.g. *Jim Baker* belongs to the class of *inner circle of advisors*. Selecting and balancing such items requires human input and cannot be automated. Because of this, Barth's work illustrates some of the problems involved in developing the use of corpora for the investigation of pragmatic features of language.

Like Aston, then, Barth uses the computer to facilitate work which could feasibly, if less conveniently, be done on paper. The corpus assists but does not drive methods of discourse analysis.

The corpus as a collection of texts II

Channell (2000) uses the Bank of English corpus to study the uses of a variety of words and phrases, such as *fat*, *right-on*, *off the beaten track, in the sticks* and *par for the course*. With respect to *par for the course*, she argues that speakers use the phrase to indicate not only that a situation is negatively evaluated, but also to confirm 'affiliation' – the sharing of experience (Channell 2000: 48–49). Using a method of examination that is not dissimilar to Aston's, she shows the detailed interaction management that is achieved by use of this phrase.

Like Aston, then, Channell uses the corpus as a collection of texts, employing search techniques to find her target phrases but then viewing the dialogue as a whole and examining it in ways that would be just as appropriate using paper and pencil as using a computer. Such a method is of course unavoidable when items with a clear discourse function, such as *thank you* or *par for the course*, are being studied. The difference between the two studies lies in corpus design. Aston uses a corpus that has been designed to investigate service encounters in order to carry out such an investigation. It would not be possible, however, to design a corpus to include instances of *par for the course*. Instead, Channell uses a corpus collected for other purposes to find examples of the item she wishes to study. If the corpus did not contain this phrase she would not study it; indeed, she

would probably use such negative evidence to argue that the phrase is not significant enough to be worth study.[3]

The corpus as a collection of words in context: qualitative

Barlow's (1996) study is concerned only with the immediate environment of each word rather than with the discourse in which it occurs (unlike Aston's study, above). Barlow examines reflexive pronouns such as *myself* and observes the most frequent verbs with which they co-occur. The verbs co-occurring most significantly with *myself* are: *FIND*, *SEE*, *TELL*, *ASK*, *LET*, *MAKE*, *STOP*, *BRING*, *GET*, *GIVE*, *CONSIDER*, *FEEL*, *ENJOY* and *HELP*. This observation suggests that *myself* is not most usefully considered as an item in a paradigm with other pronouns (Barlow notes that *I could not bring myself to watch the race* is nothing like *I could not bring him to watch the race*) but as something that occurs as an element in a number of phrases, each with its own meaning. Echoing Francis et al (1996: 62–68), Barlow also argues that the tendency to be followed by a reflexive is an important aspect of the behaviour of a number of less frequent verbs (Barlow, 1996: 12–13).

Barlow's work is a typical example of its kind. He begins with a word (although he takes a category – reflexive pronouns – he treats each lexical instance as a separate item) and examines its immediate context, comparing what is found with what intuition or grammatical theory would predict. The role of the word in the larger discourse is not commented on, except insofar as it is observable from the immediate context. As far as Barlow is concerned, then, the corpus could consist only of the word *myself* and its immediate co-texts. From the collocational information found, conclusions are drawn regarding the way that language can and should be described.

The corpus as a collection of words in context: quantitative

Mason's (1999) study also begins with a single word and ends with a theoretical conclusion, but the study is entirely quantitative and the conclusions drawn about language are statistical in nature. Here we are very far removed from the individual discourses that make up the

[3] It appears that corpora as collections are being used increasingly by linguists of various kinds. For example, Gomez-Gonzalez (1998) finds and analyses instances of extended multiple themes in a corpus, and He and Kennedy (1999) use a prosodically tagged corpus to find and categorise successful turn-bidding.

corpus in question. Like Barlow, Mason looks at the words occurring in the immediate environment of the target (node) word, but unlike Barlow he is not primarily interested in what those words are, only how many types there are in relation to the total number of concordance lines used. A low number of types means that the node word exerts a strong influence, a high number means that it exerts a weak influence. If the number of types at each point in relation to the node (one word before, two words before, one word after, two words after and so on) is calculated, it is possible to see where the node word exerts its influence. With the word *of*, for example (Mason 1999: 271), the number of types immediately before *of* is much less than that immediately after. This shows that *of* exerts more of an influence on the words it follows than on the words that follow it. Mason also shows that this influence extends to words up to four places to the left of *of*. If types of lexical items are replaced by types of word-class tag (so that all nouns count as the same thing), a similar pattern of the grammatical behaviour of a word is shown. Mason uses his study to suggest that the span used for calculating collocation (usually +/−4) should vary according to the patterning of each word.

Mason's study thus has implications for how statistics are applied to corpora. For the purposes of this interpretation, the words that make up a corpus are simply symbols, without meaning or significance. At the same time, however, Mason interprets his findings in terms of actual phrases such as *the best of*.

The corpus as a collection of categories

Aarts and Granger (1998) use a version of a group of learner corpora that is tagged to show parts of speech. The frequency of various sequences of tags in those corpora, such as 'preposition + article + noun' or 'article + adjective + noun', is then calculated. Like Mason's study, then, they treat the corpus as a collection of symbols, without meaning. In this case, the symbols are not individual words, but word-class categories. The corpora are compared in terms of the most frequent tag sequences.

Work of this kind is very abstract, but can immediately be related to specific phraseologies, as will be demonstrated in chapter 8. The abstract categories are used to produce statistical measurements, which give a starting point for less abstract explanations.

Conclusion

This chapter has briefly introduced some of the tensions that might exist when answering the question: what kind of object is a corpus? In corpus design, the criteria of size, balance and contemporaneousness are in tension with one another. In terms of use, a corpus may be constructed and/or used as a collection of individual texts, or as a collection of words in context, or as a collection of categories. Each of these implies an object of a different kind.