LOOKING AHEAD

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 A2 and A8, and further explored in Unit B1. Corpus processing (e.g. data capture, corpus mark-up and annotation) will be discussed in Units A3–A4 and A8. Using corpora in language studies will be introduced in Unit A10 and further discussed in Section B and explored in Section C of this book.

Unit A2
Representativeness, balance and sampling

A2.1 INTRODUCTION

We noted in Unit A1 that representativeness is an essential feature of a corpus. It is this feature that is typically used to distinguish a corpus from an archive (i.e. a random collection of texts). A corpus is designed to represent a particular language or language variety whereas an archive is not. Unless you are studying a dead language or highly specialized sub-language (see Unit A2.3 for further discussion), it is virtually impossible to analyse every extant utterance or sentence of a given language. Hence, sampling is unavoidable. Yet how can you be sure that the sample you are studying is representative of the language or language variety under consideration? The answer is that one must consider balance and sampling to ensure representativeness. Hence, this unit introduces the key concept of corpus representativeness as well as the related issues of balance and sampling. We will first explain what we mean by *representativeness* (Unit A2.2), followed by a discussion of the representativeness of general and specialized corpora (Unit A2.3). We will then move on to discuss corpus balance (Unit A2.4) and finally introduce sampling techniques (Unit A2.5).

A2.2 WHAT DOES REPRESENTATIVENESS MEAN IN CORPUS LINGUISTICS?

What does representativeness mean in corpus linguistics? According to Leech (1991: 27), a corpus is thought to be representative of the language variety it is supposed to represent if the findings based on its contents can be generalized to the said language variety. Biber (1993: 243) defines representativeness from the viewpoint of how this quality is achieved: 'Representativeness refers to the extent to which a sample includes the full range of variability in a population.' A corpus is essentially a *sample* of a language or language variety (i.e. *population*). Sampling is entailed in the compilation of virtually any corpus of a living language. In this respect, the representativeness of most corpora is to a great extent determined by two factors: the range of genres included in a corpus (i.e. *balance*, see Unit A2.4) and how the text chunks for each genre are selected (i.e. *sampling*, see Unit A2.5).

We noted in Unit A1.2 that the criteria used to select texts for a corpus are principally external. The external vs. internal criteria correspond to Biber's (1993: 243) situational vs. linguistic perspectives. External criteria are defined situationally irrespective of the distribution of linguistic features whereas internal criteria are defined linguistically, taking into account the distribution of such features. Biber refers to situationally defined text categories as *genres* or *registers* (see Unit A10.4), and linguistically defined text categories as *text types* (see Unit B1), though these terms are typically used interchangeably in the literature (e.g. Aston and Burnard 1998), and in this book.

Internal criteria have sometimes been proposed as a measure of corpus representativeness. Otlogetswe (2004), for example, argues that:

> The study of corpus word distributions would reveal whether words in a corpus are skewed towards certain varieties and whether in such instances it is accurate to say they are representative of the entire corpus. It would also reflect the stability of the design – whether overall representativeness is very sensitive to particular genres.

Similar views can be found elsewhere. For example, in a discussion of representativeness on the Corpora Mailing List, most discussants appeared to assume that a corpus should sufficiently represent particular words: 'A representative corpus should include the majority of the types in the language as recorded in a comprehensive dictionary' (Berber-Sardinha 1998). Such a decision would in turn entail a discussion of what should be counted as a word, e.g. whether one should count different forms of the same word as instances of the same type.

In our view, it is problematic, indeed it is circular, to use internal criteria like the distribution of words or grammatical features as the primary parameters for the selection of corpus data. A corpus is typically designed to study linguistic distributions. If the distribution of linguistic features is predetermined when the corpus is designed, there is no point in analysing such a corpus to discover naturally occurring linguistic feature distributions. The corpus has been skewed by design. As such, we generally agree with Sinclair (1995) when he says that the texts or parts of texts to be included in a corpus should be selected according to external criteria so that their linguistic characteristics are, initially at least, independent of the selection process. This view is also shared by many other scholars including Atkins, Clear and Ostler (1992: 5–6) and Biber (1993: 256). Yet once a corpus is created by using external criteria, the results of corpus analysis can be used as feedback to improve the representativeness of the corpus. In Biber's (1993: 256) words, 'the compilation of a representative corpus should proceed in a cyclical fashion'.

In addition to text selection criteria, Hunston (2002: 30) suggests that another aspect of representativeness is change over time. She claims that '[a]ny corpus that is not regularly updated rapidly becomes unrepresentative' (*ibid.*). The relevance of permanence in corpus design actually depends on how we view a corpus, i.e.

whether a corpus should be viewed as a static or dynamic language model. The static view typically applies to a *sample corpus* whereas a dynamic view applies to a *monitor corpus* (see Units A4.2 and A7.9 for further discussion). While monitor corpora following the dynamic language model are useful in tracking rapid language change such as the development and lifecycle of neologisms, they normally cover a relatively short span of time. Very long-term change can, of course, be studied using diachronic corpora such as the Helsinki Diachronic Corpus (see Units A7.7, A10.7 and B5.4), in which each component represents a specific time period. Static sample corpora, if resampled, may also allow the study of language change over time. Static sample corpora which apply the same *sampling frame* (see Unit A2.5) are particularly useful in this respect. Typical examples of this type of corpus are the Lancaster–Oslo–Bergen corpus (i.e. LOB) and the Freiburg–LOB corpus (i.e. FLOB), which represent British English in the early 1960s and the early 1990s respectively (see Unit A7.4). Another corpus following the same sampling frame is under construction on a project which is funded by the Leverhulme Trust and undertaken by Lancaster University. The corpus is designed as a match for LOB in the early 1930s. These three corpora are specifically constructed with the study of language change in mind. Diachronic corpora like the Helsinki corpus and the corpora of the LOB family are sample corpora of the static language model sort, yet they are all well suited for the study of language change.

A2.3 THE REPRESENTATIVENESS OF GENERAL AND SPECIALIZED CORPORA

There are two broad types of corpora in terms of the range of text categories represented in the corpus: *general* and *specialized* corpora. General corpora typically serve as a basis for an overall description of a language or language variety. The British National Corpus (BNC, see Unit A7.2), for example, is supposed to represent modern British English as a whole. In contrast, specialized corpora tend to be domain (e.g. medicine or law) or genre (e.g. newspaper text or academic prose) specific. For a general corpus, it is understandable that it should cover, proportionally, as many text types as possible so that the corpus is maximally representative of the language or language variety it is supposed to represent. Even a specialized corpus, e.g. one dealing with telephone calls to an operator service should be balanced by including within it a wide range of types of operator conversations (e.g. line fault, request for an engineer call-out, number check, etc.) between a range of operators and customers (see McEnery, Baker and Cheepen 2001) so that it can be claimed to represent this variety of language.

While both general and specialized corpora should be representative of a language or language variety, the representativeness of the two types of corpora are measured in different ways. The representativeness of a general corpus depends heavily on sampling from a broad range of genres (see Unit A2.4) whereas the representativeness of a specialized corpus, at the lexical level at least, can be measured by the degree of 'closure' (McEnery and Wilson 2001: 166) or 'saturation' (Belica 1996:

61–74) of the corpus. Closure/saturation for a particular linguistic feature (e.g. size of lexicon) of a variety of language (e.g. computer manuals) means that the feature appears to be finite or is subject to very limited variation beyond a certain point. To measure the saturation of a corpus, the corpus is first divided into segments of equal size based on its tokens. The corpus is said to be saturated at the lexical level if each addition of a new segment yields approximately the same number of new lexical items as the previous segment, i.e. when 'the curve of lexical growth has become asymptotic' (Teubert 1999), or is flattening out. The notion of saturation is claimed to be superior to such concepts as balance for its measurability (*ibid.*). It should be noted, however, that saturation is only concerned with lexical features. While it may be possible to adapt saturation to measure features other than lexical growth, there have been few attempts to do this to date (though see McEnery and Wilson 2001: 176–183 for a study of part-of-speech and sentence type closure).

A2.4 BALANCE

As noted in the previous section, the representativeness of a corpus, especially a general corpus, depends primarily upon how balanced the corpus is, in other words, the range of text categories included in the corpus. As with representativeness, the acceptable balance of a corpus is determined by its intended uses. Hence, a general corpus which contains both written and spoken data (e.g. the BNC, see Unit A7.2) is balanced; so are written corpora such as Brown and LOB (see Unit A7.4), and spoken corpora like CANCODE (see Unit A7.5); domain-specific corpora (e.g. the HKUST Computer Science Corpus, see Unit A7.3) can also claim to be balanced. A balanced corpus usually covers a wide range of text categories which are supposed to be representative of the language or language variety under consideration. These text categories are typically sampled proportionally (see Unit A2.5) for inclusion in a corpus so that 'it offers a manageably small scale model of the linguistic material which the corpus builders wish to study' (Atkins *et al.* 1992: 6).

While balance is often considered a *sine qua non* of corpus design, any claim of corpus balance is largely an act of faith rather than a statement of fact as, at present, there is no reliable scientific measure of corpus balance. Rather the notion relies heavily on intuition and best estimates. Nevertheless, one thing we can be certain of is that work in text typology – classifying and characterizing text categories – is highly relevant to any attempt to achieve corpus balance. Yet different ways of classifying and characterizing texts can produce different text typologies. The text typology proposed by Atkins *et al.* (1992) lists up to twenty-nine text attributes which are considered relevant in constructing a balanced corpus. All of the parameters are extra-linguistic variables, though the authors are aware that external criteria alone cannot achieve corpus balance: 'Controlling the "balance" of a corpus is something which may be undertaken only after the corpus (or at least an initial provisional corpus) has been built' (*ibid.*: 6). Yet while useful, such work is rarely the basis of corpus construction. A more typical approach to corpus balance

is that corpus-builders – for good or ill – adopt an existing corpus model when building their own corpus, assuming that balance will be achieved from the adopted model.

For example, the British National Corpus (BNC) is generally accepted as being a balanced corpus. The BNC model has been followed in the construction of a number of corpora, for example, the American National Corpus, the Korean National Corpus, the Polish National Corpus and the Russian Reference Corpus (Sharoff 2003) (see Unit A7.2 for a description of these corpora). Given the importance of such a model, a closer look at the design criteria used in building the BNC may help to give us a general idea of what is assumed to be corpus balance.

The BNC contains approximately 100 million words, of which 90 per cent are written texts and 10 per cent are transcripts of spoken data. Written texts were selected using three criteria: 'domain', 'time' and 'medium'. Domain refers to the content type (i.e. subject field) of the text; time refers to the period of text production, while medium refers to the type of text publication such as books, periodicals or unpublished manuscripts. Table A2.1 summarizes the distribution of these criteria (see Aston and Burnard 1998: 29–30). The spoken data in the BNC was collected on the basis of two criteria: 'demographic' and 'context-governed'. The demographic component is composed of informal encounters recorded by 124 volunteer respondents selected by age group, sex, social class and geographical region, while the context-governed component consists of more formal encounters such as meetings, lectures and radio broadcasts recorded in four broad context categories. The two types of spoken data complement each other, as many contexts of speech may not have been covered if demographic sampling techniques alone were used in data collection. Table A2.2 summarizes the composition of the spoken BNC. Note that in the table, the first two columns apply to both demographic and context-governed components while the third column refers to the latter component alone.

As the BNC is designed to represent contemporary British English as a whole, the overall aim of using the above text selection criteria was to achieve a balanced

Table A2.1 Composition of the written BNC

Domain	%	*Date*	%	*Medium*	%
Imaginative	21.91	1960–74	2.26	Book	58.58
Arts	8.08	1975–93	89.23	Periodical	31.08
Belief and thought	3.40	Unclassified	8.49	Misc. published	4.38
Commerce/finance	7.93			Misc. unpublished	4.00
Leisure	11.13			To-be-spoken	1.52
Natural/pure science	4.18			Unclassified	0.40
Applied science	8.21				
Social science	14.80				
World affairs	18.39				
Unclassified	1.93				

Table A2.2 Composition of the spoken BNC

Region	%	*Interaction type*	%	*Context-governed*	%
South	45.61	Monologue	18.64	Educational/informative	20.56
Midlands	23.33	Dialogue	74.87	Business	21.47
North	25.43	Unclassified	6.48	Institutional	21.86
Unclassified	5.61			Leisure	23.71
				Unclassified	12.38

selection within each text category. Aston and Burnard's (1998: 28) summary of the design criteria of the BNC illustrates the notion of corpus balance very well:

> In selecting texts for inclusion in the corpus, account was taken of both production, by sampling a wide variety of distinct types of material, and reception, by selecting instances of those types which have a wide distribution. Thus, having chosen to sample such things as popular novels, or technical writing, best-seller lists and library circulation statistics were consulted to select particular examples of them.

Balance appears to be a more important issue for a static sample corpus than for a dynamic monitor corpus. As corpora of the latter type are updated frequently, it is usually 'impossible to maintain a corpus that also includes text of many different types, as some of them are just too expensive or time consuming to collect on a regular basis' (Hunston 2002: 30–31). The builders of monitor corpora appear to feel that balance has become less of a priority – sheer size seems to have become the basis of the corpus's authority, under the implicit and arguably unwarranted assumption that a corpus will in effect balance itself when it reaches a substantial size (see Units A1.7 and A7.9 for further discussion).

Like corpus representativeness, balance is an important issue for corpus creators, corpus users and readers of corpus-based studies alike. Representativeness links to research questions. The research question one has in mind when building (or thinking of using) a corpus defines representativeness. If one wants a corpus which is representative of general English, a corpus representative of newspapers will not do. If one wants a corpus representative of newspapers, a corpus representative of *The Times* will not do. Representativeness is a fluid concept. Corpus creators should not only make their corpora as balanced as possible for the language variety in question by including a great variety of relevant representative language samples, they must also document corpus design criteria explicitly and make the documentation available to corpus users so that the latter may make appropriate claims on the basis of such corpora and decide whether or not a given corpus will allow them to pursue a specific research question. Readers of corpus-based research should also interpret the results of corpus-based studies with caution and consider whether the corpus data used in a study were appropriate. With that said, however, we entirely agree with Atkins *et al.* (1992: 6), who comment that:

> It would be short-sighted indeed to wait until one can scientifically balance a corpus before starting to use one, and hasty to dismiss the results of corpus analysis as 'unreliable' or 'irrelevant' because the corpus used cannot be proved to be 'balanced'.

A2.5 SAMPLING

Corpus representativeness and balance are closely associated with *sampling*. Given that we cannot exhaustively describe natural language, we need to sample it in order to achieve a balance and representativeness which match our research question. Having decided that sampling is inevitable, there are important decisions that must be made about how to sample so that the resulting corpus is as balanced and representative as practically possible.

As noted earlier in this unit, with few exceptions, a corpus – either a sample or monitor corpus – is typically a *sample* of a much larger *population*. A sample is assumed to be representative if what we find for the sample also holds for the general population (see Manning and Schütze 1999: 119). In the statistical sense, samples are scaled-down versions of a larger population (see Váradi 2000). The aim of sampling theory 'is to secure a sample which, subject to limitations of size, will reproduce the characteristics of the population, especially those of immediate interest, as closely as possible' (Yates 1965: 9).

In order to obtain a representative sample from a population, the first concern to be addressed is to define the *sampling unit* and the boundaries of the population. For written text, for example, a sampling unit may be a book, periodical or newspaper. The population is the assembly of all sampling units while the list of sampling units is referred to as a *sampling frame*. The population from which samples for the pioneering Brown corpus were drawn, for instance, was written English text published in the United States in 1961 while its sampling frame was a list of the collection of books and periodicals in the Brown University Library and the Providence Athenaeum. For the LOB corpus, the target population was all written English text published in the United Kingdom in 1961 while its sampling frame included the *British National Bibliography Cumulated Subject Index* 1960–1964 for books and *Willing's Press Guide* 1961 for periodicals.

In corpus design, a population can be defined in terms of language production, language reception or language as a product. The first two designs are basically demographically oriented as they use the demographic distribution (e.g. age, sex, social class) of the individuals who produce/receive language data to define the population while the last is organized around text category/genre of language data. As noted earlier, the Brown and LOB corpora were created using the criterion of language as a product while the BNC defines the population primarily on the basis of both language production and reception. However, it can be notoriously difficult to define a population or construct a sampling frame, particularly for spoken

language, for which there are no ready-made sampling frames in the form of catalogues or bibliographies.

Once the target population and the sampling frame are defined, different sampling techniques can be applied to choose a sample which is as representative as possible of the population. A basic sampling method is *simple random sampling*. With this method, all sampling units within the sampling frame are numbered and the sample is chosen by use of a table of random numbers. As the chance of an item being chosen correlates positively with its frequency in the population, simple random sampling may generate a sample that does not include relatively rare items in the population, even though they can be of interest to researchers. One solution to this problem is *stratified random sampling*, which first divides the whole population into relatively homogeneous groups (so-called *strata*) and samples each stratum at random. In the Brown and LOB corpora, for example, the target population for each corpus was first grouped into fifteen text categories such as news reportage, academic prose and different types of fiction; samples were then drawn from each text category. Demographic sampling, which first categorizes sampling units in the population on the basis of speaker/writer age, sex and social class, is also a type of stratified sampling. Biber (1993) observes that a stratified sample is never less representative than a simple random sample.

A further decision to be made in sampling relates to sample size. For example, with written language, should we sample full texts (i.e. whole documents) or text chunks? If text chunks are to be sampled, should we sample text initial, middle or end chunks? Full text samples are certainly useful in text linguistics, yet they may potentially constitute a challenge in dealing with vexatious copyright issues (see Unit A9). Also, given its finite overall size, the coverage of a corpus including full texts may not be as balanced as a corpus including text segments of constant size, and 'the peculiarity of an individual style or topic may occasionally show through into the generalities' (Sinclair 1991a: 19). Aston and Burnard (1998: 22) argue that the notion of 'completeness' may sometimes be 'inappropriate or problematic'. As such, unless a corpus is created to study such features as textual organization, or copyright holders have granted you permission to use full texts, it is advisable to sample text segments. According to Biber (1993: 252), frequent linguistic features are quite stable in their distributions and hence short text chunks (e.g. 2,000 running words) are usually sufficient for the study of such features while rare features are more varied in their distribution and thus require larger samples. In selecting samples to be included in a corpus, however, attention must also be paid to ensure that text initial, middle and end samples are balanced.

Another sampling issue, which particularly relates to stratified sampling, is the proportion and number of samples for each text category. The numbers of samples across text categories should be proportional to their frequencies and/or weights in the target population in order for the resulting corpus to be considered as representative. Nevertheless, it has been observed that, as with defining a target population, such proportions can be difficult to determine objectively (see Hunston

2002: 28–30). Furthermore, the criteria used to classify texts into different categories or genres are often dependent on intuitions. As such, the representativeness of a corpus, as noted, should be viewed as a statement of belief rather than fact. In the Brown corpus, for example, the ratios between the fifteen text categories were determined by a panel of experts (see Table A7.1, p. 62). As for the number of samples required for each category, Biber (1993) demonstrates that ten 2,000-word samples are typically sufficient.

The above discussion suggests that in constructing a balanced, representative corpus, stratified random sampling is to be preferred over simple random sampling while different sampling methods should be used to select different types of data. For written texts, a text typology established on the basis of external criteria is highly relevant while for spoken data demographic sampling is appropriate. However, samples obtained from demographic sampling must be complemented by context-governed sampling so that some contextually governed linguistic variations can be included in the resulting corpus.

Summary

This unit introduced some important concepts in corpus linguistics – representativeness, balance and sampling. A corpus is considered representative if what we find on the basis of the corpus also holds for the language or language variety it is supposed to represent. For most corpora, representativeness is typically achieved by balancing, i.e. covering a wide variety of frequent and important text categories that are proportionally sampled from the target population. Claims of corpus representativeness and balance, however, should be interpreted in relative terms and considered as a statement of faith rather than as fact, as presently there is no objective way to balance a corpus or to measure its representativeness. Furthermore, it is only by considering the research question one has to address that one is able to determine what is an acceptable balance for the corpus one should use and whether it is suitably representative. The concepts introduced in this unit will help you to determine if a particular corpus is suitable for your intended research. They are also helpful in determining whether a research question is amenable to corpus analysis.

LOOKING AHEAD

The notions of corpus balance and representativeness will be discussed further in Units A8.3 and B1, while the potential uses of corpora in language studies will be explored in Unit A10. Units A7.9 and B2 will further develop some issues touched upon in this unit such as the monitor corpus model and the pros and cons of the corpus-based approach. In the following two units, we will introduce two further concepts in corpus linguistics, namely *mark-up* and *annotation*.