Unit 11 Corpus representativeness and balance

11.1 Introduction

We learnt from units 1 and 2 in Section A that one of the commonly accepted defining features of a corpus is representativeness. Representativeness is typically achieved by balancing the corpus through sampling a wide range of text categories which are defined primarily in terms of external criteria. It was also noted that it could be difficult both to define a target population and to determine the proportions across text categories. In this unit, we discuss corpus representativeness and balance, using two excerpts from published papers. This discussion will provide a more thorough grounding in these ideas than has been achieved so far in the book.

11.2 Biber (1993)

Biber has published widely on the issue of corpus design. In this section we present an extract from his paper 'Representativeness in corpus design', originally published in *Literary and Linguistic Computing* in 1993. In this paper, Biber addresses a number of issues related to how to achieve corpus representativeness, including the meaning of representativeness, defining a target population, stratified vs. proportional sampling, sampling within texts, and issues relating to sample size. Biber's ideas of corpus representativeness are generally accepted and certainly widely reported (e.g. McEnery and Wilson 2001; Tognini-Bonelli 2001; Hunston 2002). The extract below is from the first section of the paper.

Biber, D. 1993. 'Representativeness in corpus design'. *Literary and Linguistic Computing* 8/4: 243-57.

Some of the first considerations in constructing a corpus concern the overall design: for example, the kinds of texts included, the number of texts, the selection of particular texts, the selection of text samples from within texts, and the length of text samples. Each of these involves a sampling decision, either conscious or not.

The use of computer-based corpora provides a solid empirical foundation for general purpose language tools and descriptions, and enables analyses of a scope not otherwise possible. However, a corpus must be 'representative' in order to be appropriately used as the basis for generalizations concerning a language as a whole; for example, corpus-based dictionaries, grammars, and general part-of-speech taggers are applications requiring a representative basis (cf. Biber, 1993b).

Typically researchers focus on sample size as the most important consideration in achieving representativeness: how many texts must be included in the corpus, and how many words per text sample. Books on sampling theory, however, emphasize that sample size is not the most important consideration in selecting a representative sample; rather, a thorough definition of the target population and decisions concerning the method of sampling are prior considerations. Representativeness refers to the extent to which a sample includes the full range of variability in a population. In corpus design, variability can be considered from situational and from linguistic perspectives, and both of these are important in determining representativeness. Thus a corpus design can be evaluated for the extent to which it includes: (1) the range of text types in a language, and (2) the range of linguistic distributions in a language.

Any selection of texts is a sample. Whether or not a sample is 'representative', however, depends first of all on the extent to which it is selected from the range of text types in the

target population; an assessment of this representativeness thus depends on a prior full definition of the 'population' that the sample is intended to represent, and the techniques used to select the sample from that population. Definition of the target population has at least two aspects: (1) the boundaries of the population — what texts are included and excluded from the population; (2) hierarchical organization within the population — what text categories are included in the population, and what are their definitions. In designing text corpora, these concerns are often not given sufficient attention, and samples are collected without a prior definition of the target population. As a result, there is no possible way to evaluate the adequacy or representativeness of such a corpus (because there is no well-defined conception of what the sample is intended to represent).

In addition, the representativeness of a corpus depends on the extent to which it includes the range of linguistic distributions in the population; i.e. different linguistic features are differently distributed (within texts, across texts, across text types), and a representative corpus must enable analysis of these various distributions. This condition of linguistic representativeness depends on the first condition; i.e. if a corpus does not represent the range of text types in a population, it will not represent the range of linguistic distributions. In addition, linguistic representativeness depends on issues such as the number of words per text sample, the number of samples per 'text', and the number of texts per text type. These issues will be addressed in Sections 3 and 4.

However, the issue of population definition is the first concern in corpus design. To illustrate, consider the population definitions underlying the Brown corpus (Francis and Kucera 1964/79) and the LOB corpus (Johansson *et al.*, 1978). These target populations were defined both with respect to their boundaries (all published English texts printed in 1961, in the United States and United Kingdom respectively), and their hierarchical organizations (fifteen major text categories and numerous subgenre distinctions within these categories). In constructing these corpora, the compilers also had good 'sampling frames', enabling probabilistic, random sampling of the population. A sampling frame is an operational definition of the population, an itemized listing of population members from which a representative sample can be chosen. The LOB corpus manual (Johansson *et al.*, 1978) is fairly explicit about the sampling frame used: for books, the target population was operationalized as all 1961 publications listed in *The British National Bibliography Cumulated Subject Index, 1960–1964* (which is based on the subject divisions of the Dewey Decimal Classification system), and for periodicals and newspapers, the target population was operationalized as all 1961 publications listed in *Willing's Press Guide* (1961). In the case of the Brown corpus, the sampling frame was the collection of books and periodicals in the Brown University Library and the Providence Athenaeum; this sampling frame is less representative of the total texts in print in 1961 than the frames used for construction of the Lancaster-Oslo/Bergen (LOB) corpus, but it provided well-defined boundaries and an itemized listing of members. In choosing and evaluating a sampling frame, considerations of efficiency and cost effectiveness must be balanced against higher degrees of representativeness.

Given an adequate sampling frame, it is possible to select a probabilistic sample. There are several kinds of probabilistic samples, but they all rely on random selection. In a simple random sampling, all texts in the population have an equal chance of being selected. For example, if all entries in the *British National Bibliography* were numbered sequentially, then a table of random numbers could be used to select a random sample of books. Another method of probabilistic sampling, which was apparently used in the construction of the Brown and LOB corpora, is 'stratified sampling'. In this method, subgroups are identified within the target population (in this case, the genres), and then each of those 'strata' are sampled using random techniques. This approach has the advantage of guaranteeing that all strata are adequately represented while at the same time selecting a non-biased sample within each stratum (i.e. in the case of the Brown and LOB corpora, there was 100% representation at the level of genre categories and an unbiased selection of texts within each genre).

Note that, for two reasons, a careful definition and analysis of the non-linguistic characteristics of the target population is a crucial prerequisite to sampling decisions. First, it is not possible to identify an adequate sampling frame or to evaluate the extent to which a particular sample represents a population until the population itself has been carefully defined. A good illustration is a corpus intended to represent the spoken texts in a language. As there are no catalogues or bibliographies of spoken texts, and since we are all constantly expanding the universe of spoken texts in our everyday conversations, identifying an adequate sampling frame in this case is difficult: but without a prior definition of the boundaries and parameters of speech within a language, evaluation of a given sample is not possible.

The second motivation for a prior definition of the population is that stratified samples are almost always more representative than non-stratified samples (and they are nevertheless representative). This is because identified strata can be fully represented (100% sampling) in the proportions desired, rather than depending on random selection techniques. In statistical terms, the between-group variance is typically larger than within-group variance and thus a sample that forces representation across identifiable groups will be more representative overall. Returning to the Brown and LOB corpora, a prior identification of the genre categories (e.g. press reportage, academic prose, and mystery fiction) and subgenre categories (e.g. medicine, mathematics, and humanities within the genre of academic prose) guaranteed 100% representation at those two levels; i.e. the corpus builders attempted to compile an exhaustive listing of the major text categories of published English prose, and all of these categories were included in the corpus design. Therefore, random sampling techniques were required only to obtain a representative selection of texts from within each subgenre. The alternative, a random selection from the universe of all published texts, would depend on a large sample and the probabilities associated with random selection to assure representation of the range of variation at all levels (across genres, subgenres, and texts within subgenres), a more difficult task.

11.3 Atkins, Clear and Ostler (1992)

The excerpt included in this section is extracted from Atkins et al's paper 'Corpus design criteria', originally published in *Literary and Linguistic Computing* in 1992. This excerpt (section 4 of the paper) addresses the major difficulties in defining a target population, contrasting the sets of texts received vs. those produced by a target group, and the internal (linguistic) vs. external (social) means of defining such groups.

Atkins, S., Clear, J. and Ostler, N. 1992. 'Corpus design criteria'. *Literary and Linguistic Computing* 7/1: 1-16.

4. Population and Sampling

In building a natural language corpus one would like ideally to adhere to the theoretical principles of statistic sampling and inference. Unfortunately, the standard approaches to statistical sampling are hardly applicable to building a language corpus. First, it is very difficult (often impossible) to delimit the total population in any rigorous way. Textbooks on statistical methods almost always focus on clearly defined populations. Secondly, even if the population could be delimited, because of the sheer size of the population and given current and foreseeable resources, it will always be possible to demonstrate that some feature of the population is not adequately represented in the sample. Thirdly, there is no obvious unit of language (words? sentences? texts?) which is to be sampled and which can be used to define the population. We may sample words or sentences or 'texts' among other things. Despite these difficulties, some practical basis for progress can be established. An approach suggested by Woods, Fletcher, and Hughes is to accept the results of each study as though any sampling had been carried out in the theoretically 'correct' way, to attempt to foresee possible objections. In corpus linguistics such a pragmatic approach seems the only course of action. Moreover, there is a tendency to overstate the possibility and effects of experimental error:

indeed, good scientific estimation of the possibility and scale of experimental error in statistics of natural language corpora is seldom carried out at all.

All samples are *biased* in some way. Indeed the sampling problem is precisely that a corpus is inevitably biased in some respects. The corpus users must continually evaluate the results drawn from their studies and should be encouraged to report them (see Subsection 2.5).

The difficulty of drawing firm conclusions when the number of observed instances is few underlines the methodological point made by Woods, Fletcher, and Hughes: that researchers should question how the sample was obtained and assess whether this is likely to have a bearing on the validity of the conclusions reached.

4.1 Defining the Population

When a corpus is being set up as a sample with the intention that observation of the sample will allow us to make generalizations about language, then the relationship between the sample and the target population is very important. The more highly specialized the language to be sampled in the corpus, the fewer will be the problems in defining the texts to be sampled. For a general-language corpus, however, there is a primary decision to be made about whether to sample the language that people hear and read (their *reception*) or the language that they speak and write (their *production*).

Defining the population in terms of language reception assigns tremendous weight to a tiny proportion of the writers and speakers whose language output is received by a very wide audience through the media. However, most linguists would reject the suggestion that the language of the daily tabloid newspapers (though they may have a very wide reception) can be taken to represent the language production of any individual member of the speech community.

The corpus builder has to remain aware of the reception and production aspects, and though texts which have a wide reception are by definition easier to come by, if the corpus is to be a true reflection of native speaker usage, then every effort must be made to include as much production material as possible. For a large proportion of the language community, writing (certainly any extended composition) is a rare language activity. Judged on either of these scales, private conversation merits inclusion as a significant component of a representative general language corpus. Judged in terms of production, personal and business correspondence and other informal written communications form a valuable contribution to the corpus.

To summarize, we can define the language to be sampled in terms of language production (many producers each with few receivers) and language reception (few producers but each with many receivers). Production is likely to be greatly influenced by reception, but technically only production defines the language variety under investigation. However, collection of a representative sample of total language production is not feasible. The compiler of a general language corpus will have to evaluate text samples on the basis of *both* reception and production.

4.2 Describing the Population

A distinction between external and internal criteria is of particular importance for constructing a corpus for linguistic analysis. The internal criteria are those which are essentially *linguistic*: for example, to classify a text as formal/informal is to classify it according to its linguistic characteristics (lexis/diction and syntax). External criteria are those which are essentially *non-linguistic*. Section 6 contains a list of attributes which we consider relevant to the description of the language population from which corpus texts are to be sampled. These attributes, however, are all founded upon extra-linguistic features of texts (external evidence). Of course, the internal criteria are not independent of the external ones and the interrelation between them is one of the areas of study for which a corpus is of primary value. In general, external criteria can be determined without reading the text in question, thereby ensuring that no linguistic judgements are being made. The initial selection of texts for inclusion in a corpus will inevitably be based on external evidence primarily. Once the text is captured and subject to analysis there will be a range of linguistic features of the text which will contribute to its characterization in terms of internal evidence. A corpus

selected entirely on internal criteria would yield no information about the relation between language and its context of situation. A corpus selected entirely on external criteria would be liable to miss significant variation among texts since its categories are not motivated by textual (but by contextual) factors.

11.4 Unit summary and looking ahead

This unit discussed in more detail the key concepts of corpus representativeness and balance as introduced in unit 2. It is clear that in order to achieve corpus balance and representativeness, it is essential to define the target population and apply appropriate sampling techniques. There is also a consensus that external (or situational, social or extra-linguistic) rather than internal (or linguistic) criteria should be used in initial corpus design. It is important to note that corpus representativeness and balance are also closely associated with the sample vs. monitor corpus models. Readers are advised to refer to units 2.3 and 7.9 for a discussion of this debate. In the next unit, we will discuss the pros and cons of the corpus-based approach.