3.2. QUALITATIVE VS. QUANTITATIVE ANALYSIS

The difference between qualitative and quantitative corpus analysis, as the terms themselves imply, is that in qualitative research no attempt is made to assign frequencies to the linguistic features which are identified in the data. Whereas in quantitative research we classify features, count them and even construct more complex statistical models in an attempt to explain what is observed, in qualitative research the data are used only as a basis for identifying and describing aspects of usage in the language and to provide 'real-life' examples of particular phenomena.

As Schmied (1993) has observed, a stage of qualitative research is often a precursor for quantitative analysis, since, before linguistic phenomena are classified and counted, the categories for classification must first be identified.[1] But it is more useful to consider these two as forming two different, but not necessarily incompatible, perspectives on corpus data.

Qualitative forms of analysis offer a rich and detailed perspective on the data. In qualitative analyses, rare phenomena receive, or at least ought to receive, the same attention as more frequent phenomena and, because the aim is complete detailed description rather than quantification, delicate variation in the data is foregrounded: qualitative analysis enables very fine distinctions to be drawn since it is not necessary to shoehorn the data into a finite number of classifications. The fact that qualitative analysis is not primarily classificatory also means that the ambiguity which is inherent in human language – not only by accident but also through the deliberate intent of language users – can be fully recognised in the analysis: qualitative research does not force a potentially misleading intepretation. For instance, in a quantitative stylistic analysis it might be necessary to classify the word *red* as either simply a colour or as a political categorisation (signifying socialism or communism): in a qualitative analysis both the senses of *red* in a phrase such as *the red flag* could be recognised – the physical property of the flag's colour and its political significance. However, the main disadvantage of qualitative approaches to corpus analysis is that their findings cannot be extended to wider populations with the same degree of certainty with which quantitative analyses can, because, although the corpus may be statistically representative, the specific findings of the research cannot be tested to discover whether they are statistically significant or more likely to be due to chance.

In contrast to qualitative analysis, the quantitiative analysis of a sampled corpus does allow for its findings to be generalised to a larger population and, furthermore, it means that direct comparisons may be made between different corpora, at least so long as valid sampling and significance techniques have been employed. Quantitative analysis thus enables one to separate the wheat from the chaff: it enables one to discover which phenomena are likely to be genuine reflections of the behaviour of a language or variety and which are merely chance occurrences. In the more basic task of looking non-comparatively at a single language variety, quantitative analysis enables one to get a precise picture of the frequency and rarity of particular phenomena and hence, arguably, of their relative normality or abnormality. However, the picture of the data which emerges from quantitative analysis is necessarily less rich than that obtained from qualitative analysis. Quantification, as suggested above, entails classification. For statistical purposes, these classifications have to be of the hard-and-fast (so-called 'Aristotelian') type, that is an item either belongs in class *x* or it doesn't: to take our example of *red* again, we would have to decide firmly whether to put the word in the category 'colour' or the category 'politics'. In practice, however, many linguistic items and phenomena do not fit this Aristotelian model: rather, they are consistent with the more recent notions of 'fuzzy sets', where some phenomena may clearly belong in class *x* but others have a more dubious status and may belong in potentially more than one class, as was the case above with the word *red*. Quantitative analysis may therefore entail in some circumstances a certain idealisation of the data: it forces the analyst to make a decision which is perhaps not a 100 per cent accurate reflection of the reality contained in the data. At the same time, quantitative analysis also tends to sideline rare occurrences. To ensure that certain statistical significance tests (such as the chi-squared test which we shall meet later in section 3.4.3) provide reliable results, it is essential that specific minimum frequencies are obtained and this can mean that fine distinctions have to be deliberately blurred to ensure that statistical significances can be computed, with a resulting loss of data richness.

It will be appreciated from this brief discussion that both qualitative and quantitative analyses have something to contribute to corpus study. Qualitative analysis can provide greater richness and precision, whereas quantitative analysis can provide statistically reliable and generalisable results. There has recently been a move in social science research towards multi-method approaches which largely reject the narrow analytical paradigms in favour of the breadth of information which the use of more than one method may provide. Corpus linguistics could, as Schmied (1993) demonstrates, benefit as much as any field from such multi-method research, combining both qualitative and quantitative perspectives on the same phenomena.

3.3. CORPUS REPRESENTATIVENESS

Although quantitative analyses may be carried out on any sample of text, these can be misleading if one wants to generalise the findings on that sample to some larger population, for example a genre as a whole. This, as we have already seen in Chapter 1, was essentially the foundation of Chomsky's criticism of early corpus linguistics. Chomsky took the view that because a corpus *did* constitute only a small sample of a large and potentially infinite population – namely the set of possible sentences of a language – it would be skewed and hence unrepresentative of the population as a whole. This is a valid criticism –

it is true with any kind of sample that rare elements *may* occur in higher proportions and frequent elements in lesser proportions than in the population as a whole – and this criticism applies not only to linguistic corpora but to any form of scientific investigation which is based on sampling rather than on the exhaustive analysis of an entire and finite population: in other words, it applies to a very large proportion of the scientific and social scientific research which is carried out today. However, the effects of Chomsky's criticism are not quite so drastic as it appears at first glance, since there are many safeguards which may be applied in sampling for maximal representativeness.

The reader will recall from Chapter 1 that, at the time when Chomsky first made his criticism in the 1950s, most corpora were very small entities. This was due as much to necessity as to choice: the development of text analysis by computer had still to progress considerably and thus corpora had still largely to be analysed by hand. Hence these corpora had to be of a manageable size for manual analysis. Although size – short of including the whole target population – is not a guarantee of representativeness, it does enter significantly into the factors and calculations which need to be considered in producing a maximally representative corpus. Small corpora tend only to be representative for certain high frequency linguistic features, and thus Chomsky's criticism was at least partly true of these early corpora. But since today we have powerful computers which can readily store, search and manipulate many millions of words, the issue of size is no longer such a problem and we can attempt to make much more representative corpora than Chomsky could dream of when he first criticised corpus-based linguistics.

In discussing the ways of achieving the maximal degree representativeness, it should first be emphasised once again that in producing a corpus we are dealing with a **sample** of a much larger **population**. Random sampling techniques in themselves are standard to many areas of science and social science, and these same techniques are also used in corpus building. But there are particular additional caveats which the corpus builder must be aware of.

Biber (1993b), in a detailed survey of this issue, emphasises as the first step in corpus sampling the need to define as clearly as possible the limits of the population which we are aiming to study before we can proceed to define sampling procedures for it. This means that we should not start off by saying vaguely that we are interested in, for instance, the written German of 1993, but that we must actually rigorously define what the boundaries of 'the written German of 1993' are for our present purpose, that is, what our **sampling frame** – the entire population of texts from which we will take our samples – is. Two approaches have been taken to this question in the building of corpora of written language. The first approach is to use a comprehensive bibliographical index. So, for 'the written German of 1993', we might define our sampling frame as being the entire contents of an index of published works in German for that year, for example, the *Deutsche National-Bibliographie*. This is the

approach which was taken by the Lancaster-Oslo/Bergen corpus, using the *British National Bibliography* and *Willings' Press Guide* as the indices. The second possible approach is to define the sampling frame as being the holdings of a given library which belong to the variety and period in which we are interested. So, for our example, we might define the sampling frame as being all the German-language books and periodicals in Lancaster University Library which were published in 1993. This latter approach is the one which was taken in building the Brown corpus and also the Guangzhou Petroleum English Corpus.

These approaches are all well and good with published works such as books or newspapers, but it is not possible to use them with informal language such as conversations or private correspondence, since such kinds of language are not formally indexed or stored in a library. In these cases, therefore, instead of basing the sampling frame on an index it is usual to employ demographic sampling of the kind which will be familiar from its use in public opinion research, that is, selecting informants on the basis of their age, sex, region, social class and so on. This is a method which was used in collecting the spoken parts of the British National Corpus: informants were selected on the basis of demographic sampling and were then given personal-stereo cassette recorders on which they recorded their everyday spoken interactions for a period of two to seven days (Crowdy 1993). However, as Crowdy notes, this kind of demographic sampling can miss out on many important language types and hence it is often necessary to supplement the demographic sampling with a more context governed approach. Again, this is what was done in the BNC project. It was recognised that such important spoken activities as broadcast interviews and legal proceedings would probably not enter into the interactions recorded by the informants and so a selection of contextually determined linguistic activity types such as these were defined and sampled in addition to the demographically sampled corpus; the former make up approximately half of the entire spoken corpus.

In addition to defining the population itself, Biber (1993b) also emphasises the advantage of determining beforehand the hierarchical structure (or **strata**) of the population, that is, defining what different genres, channels and so on it is made up of. So, going back to our example of written German, we could say it is made up of genres such as newspaper reporting, romantic fiction, legal statutes, scientific writing and so on. Biber observes that stratificational sampling is never less representative than pure probabilistic sampling and is often more so, since it enables each individual stratum to be subjected to a full probabilistic sampling. But, to state the opposite case, it has to be said that these strata, like corpus annotation, are an act of interpretation on the part of the corpus builder because they are founded on particular ways of dividing up language into entities such as genres which, it may be argued, are not naturally inherent within it: different linguists may specify different genre groupings

according to their theoretical perspectives on linguistic variation.

Having defined the population, one needs to determine which sample sizes are most representative of it, both in terms of the optimal *length* of each sample text and the optimal *number* of texts which should be included in the corpus. Both these figures are ultimately dependent on the distribution of linguistic features within the population, that is, what is the probability that y text samples of length n will contain proportionately the same number and distribution of examples of particular items as the total population? In a pilot study, Biber found that frequent items are stable in their distributions and hence small samples are adequate for these. Rarer features on the other hand show more variation in their distributions and consequently require larger samples if they are to be fully represented in the corpus, as de Haan (1992) has also observed. In terms of such rarer features, therefore, we can perhaps admit that Chomsky's criticism of the small corpora of the 1950s was a valid one.

Biber notes that the standard statistical equations which are used to determine these optimal sample lengths and sample numbers are problematic for corpus building (1993b). This is because they require two statistical values which cannot be computed for a corpus as a whole: **standard deviations**, which must be calculated for each individual feature, and **tolerable error**, which will vary according to the overall frequency of a feature. These values are, therefore, problematic, since a corpus, unless collected for one specific purpose, is normally intended for use in research on many different features of language. Biber's suggestion in this situation is that the most conservative way of ensuring representative samples is to base the computations on the most widely varying feature. With regard to sample lengths, taking samples of sizes which are representative of that feature should mean that the samples are also representative of those features which show less variation in distribution. Similarly, with the number of texts within each genre, the degree of variation on that feature which occurs within given genres is used to scale the number of texts required to represent each genre.

It will be appreciated, then, that corpus sampling is by no means a straightforward exercise. However, the constant application of strict statistical procedures should ensure that the corpus is as representative as possible of the larger population, within the limits imposed by practicality.

One way of supplementing these procedures for enhancing the representativeness of corpora is the use of dispersion statistics. Dispersion is a measure of how evenly distributed the occurrence of a feature is in a text or corpus – for example, whether its appearance is restricted mainly to a few places or whether it occurs much more widely. Frequency alone cannot be a measure of typicality: in a corpus of ten genres, two words might both have a frequency of 20, but one of these words might have two occurrences in each of the ten genres whereas the other's 20 occurrences might all be concentrated within a single genre. Dispersion can thus tell us how typical a word is and not just how often

it occurs: it can serve to counterbalance the concern, voiced amongst others by Chomsky, about the potential skewedness of corpora. To take a couple of fairly obvious examples, we might expect the words 'the' and 'and' to be very widely, and quite evenly, distributed, whereas a word such as 'autopsy' would occur only rarely outside certain text types – e.g. crime reporting, crime fiction and medical writing – but might occur very frequently within those text types. We can also use dispersion to examine the distribution of different senses of the same word – i.e. which are the most typical and which are more specialised.

The most reliable dispersion measure has been found to be Juilland's D coefficient (Lyne 1985). For this equation and a brief discussion of it in the context of other dispersion measures, see the volume by Oakes in this series (Oakes 1998: 189–92).

Dispersion measures, though used in early computer-based work on word frequencies (see the discussion of Juilland's work in Chapter 1), have been rather neglected in modern corpus linguistics, but have in the last few years come to prominence again – see, for example, the study of fixed expressions by De Cock *et al.* (1998).

3.4. APPROACHING QUANTITATIVE DATA

In the preceding sections, we have seen the value of supplementing qualitative analyses of language with quantitative data. We have also seen why corpora in particular are of value for quantitative linguistic analysis. But it should be noted that the use of quantification in corpus linguistics typically goes well beyond simple counting: many sophisticated statistical techniques are used which can both provide a mathematically rigorous analysis of often complex data – one might almost say, colloquially, to bring order out of chaos – and be used to show with some degree of certainty that differences between texts, genres, languages and so on are real ones and not simply a fluke of the sampling procedure.

In this section we introduce briefly some of the quantitative methods which are of most value in working practically with corpora. But before we move on we must raise two notes of caution. First, this section is of necessity incomplete. There are very many statistical techniques which have been, or can potentially be, applied to corpora and space precludes coverage of all of them. Instead we have chosen to concentrate on those which we consider to be the most important and most widely used. Second, we do not aim here to provide a complete step-by-step 'how to do it' guide to statistics. Many of the statistical techniques used in corpus linguistics are very complex and most require the use of computer software for them to be made manageable. To explain the mathematics fully we would need something approaching a full chapter for each technique. What we have done instead is to try to outline with as little mathematics as possible what each technique does and why it is of practical value to the corpus linguist. Other books in the Edinburgh Textbooks in