25

REPRESENTATIVENESS IN CORPUS DESIGN

Douglas Biber

Source: *Literary and Linguistic Computing*, 8:4 (1993), 243–257.

Abstract

The present paper addresses a number of issues related to achieving 'representativeness' in linguistic corpus design, including: discussion of what it means to 'represent' a language, definition of the target population, stratified versus proportional sampling of a language, sampling within texts, and issues relating to the required sample size (number of texts) of a corpus. The paper distinguishes among various ways that linguistic features can be distributed within and across texts; it analyses the distributions of several particular features, and it discusses the implications of these distributions for corpus design.

The paper argues that theoretical research should be prior in corpus design, to identify the situational parameters that distinguish among texts in a speech community, and to identify the types of linguistic features that will be analysed in the corpus. These theoretical considerations should be complemented by empirical investigations of linguistic variation in a pilot corpus of texts, as a basis for specific sampling decisions. The actual construction of a corpus would then proceed in cycles: the original design based on theoretical and pilot-study analyses, followed by collection of texts, followed by further empirical investigations of linguistic variation and revision of the design.

1 General considerations

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, 1960–1904* (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 never less 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.[1] 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.

In the present paper, I will first consider issues relating to population definitions for language corpora and attempt to develop a framework for stratified analysis of the corpus population (Section 2). In Section 3, then, I will return to particular sampling issues, including proportional versus non-proportional sampling, sampling within texts (how many words per text and stratified sampling within texts), and issues relating to sample size. In Section 4, I will describe differences in the distributions of linguistic features, presenting the distributions of several particular features, and I will discuss the implications of these distributions for corpus design. Finally, in Section 5, I offer a brief overview of corpus design in practice.

2 Strata in a text corpus: an operational proposal concerning the salient parameters of register and dialect variation

As noted in the last section, definition of the corpus population requires specification of the boundaries and specification of the strata. If we adopt the ambitious goal of representing a complete language, the population boundaries can be specified as all of the texts in the language. Specifying the relevant strata and identifying sampling frames are obviously more difficult tasks, requiring a theoretically motivated and complete specification of the kinds of texts. In the present section I offer a preliminary proposal for identifying the strata for such a corpus and operationalizing them as sampling frames. The proposal is restricted to western societies (with examples from the United States), and is intended primarily as an illustration rather than a final solution, showing how a corpus of this kind could be designed.

I use the terms genre or register to refer to situationally defined text categories (such as fiction, sports broadcasts, psychology articles), and text type to refer to linguistically defined text categories. Both of these text classification systems are valid, but they have different bases. Although registers/genres are not defined on linguistic grounds, there are statistically important linguistic differences among these categories (Biber, 1986, 1988), and linguistic feature counts are relatively stable across texts within a register (Biber, 1990). In contrast, text types are identifed on the basis of shared linguistic co-occurrence patterns, so that the texts within each type are maximally similar in their linguistic characteristics, while the different types are maximally distinct from one another (Biber, 1989).

In defining the population for a corpus, register/genre distinctions take precedence over text type distinctions. This is because registers are based on criteria external to the corpus, while text types are based on internal criteria: i.e. registers are based on the different situations, purposes, and functions of text in a speech community, and these can be identifed prior to the construction of a corpus. In contrast, identification of the salient text type distinctions in a language requires a representative corpus of texts for analysis; there is no a priori way to identify linguistically defined types. As I show in Section 4, though, the results of previous studies, as well as on-going research during the construction of a corpus, can be used to assure that the selection of texts is linguistically as well as situationally representative.

For the most part, corpus linguistics has concentrated on register differences.[2] In planning the design of a corpus, however, decisions must be made whether to include a representative range of dialects or to restrict the corpus to a single dialect (e.g. a 'standard' variety). Dialect parameters specify the demographic characteristics of the speakers and writers, including geographic region, age, sex, social class, ethnic group, education, and occupation.[3]

Different overall corpus designs represent different populations and meet different research purposes. Three of the possible overall designs are

organized around text production, text reception, and texts as products. The first two of these are demographically organized at the top level; i.e. individuals are selected from a larger demographic population, and then these individuals are tracked to record their language use. A production design would include the texts (spoken and written) actually produced by the individuals in the sample; a reception design would include the texts listened to or read. These two approaches would address the question of what people actually do with language on a regular basis. The demographic selection could be stratified along the lines of occupation, sex, age, etc.

A demographically oriented corpus would not represent the range of text types in a language, since many kinds of language are rarely used, even though they are important on other grounds. For example, few individuals will ever write a law or treaty, an insurance contract, or a book of any kind, and some of these kinds of texts are also rarely read. It would thus be difficult to stratify a demographic corpus in such a way that it would insure representativeness of the range of text categories. Many of these categories are very important, however, in defining a culture. A corpus organized around texts as products would be designed to represent the range of registers and text types rather than the typical patterns of use of various demographic groups.

Work on the parameters of register variation has been carried out by anthropological linguists such as Hymes and Duranti, and by functional linguists such as Halliday (see Hymes 1974; Brown and Fraser, 1979; Duranti, 1985; Halliday and Hasan, 1989). In Biber (1993a), I attempt to develop a relatively complete framework, arguing that 'register' should be specified as a continuous (rather than discrete) notion, and distinguishing among the range of situational differences that have been considered in register studies. This framework is overspecified for corpus design work—values on some parameters are entailed by values on other parameters, and some parameters arc specific to restricted kinds of texts. Attempting to sample at this level of specificity would thus be extremely difficult. For this reason I propose in Table 1 a reduced set of sampling strata, balancing operational feasibility with the desire to define the target population as completely as possible.

The first of the above parameters divides the corpus into three major components: writing, speech, and scripted speech. Each of these three requires different sampling considerations, and thus not all subsequent situational parameters are relevant for each component.

Within writing, the first important distinction is publication.[4] This is because the population of published texts can be operationally bounded, and various catalogues and indexes provide itemized listings of members. For example, the following criteria might be used for the operational definition of 'published' texts: (1) they are printed in multiple copies for distribution; (2) they are copyright registered or recorded by a major indexing service. In the United States, a record of all copyright registered books and periodicals

Table 1 Situational parameters listed as hierarchical sampling strata.

1	*Primary channel.* Written/spoken/scripted speech
2	*Format.* Published/not published (+ various formats within 'published')
3	*Setting.* Institutional/other public/private-personal
4	*Addressee.*
	(*a*) Plurality. Unenumerated/plural/individual/self
	(*b*) Presence (place and time). Present/absent
	(*c*) Interactiveness. None/little/extensive
	(*d*) Shared knowledge. General/specialized/personal
5	*Addressor.*
	(*a*) *Demographic variation.* Sex, age, occupation, etc.
	(*b*) *Acknowledgement.* Acknowledged individual/institution
6	*Factuality.* Factual-informational/intermediate or indeterminate/imaginative
7	*Purposes.* Persuade, entertain, edify, inform, instruct, explain, narrate, describe, keep records, reveal self, express attitudes, opinions, or emotions, enhance interpersonal relationship, . . .
8	*Topics.* . . .

is available at the Library of Congress. Other 'published' texts that are not copyright registered include government reports and documents, legal reports and documents, certain magazines and newspapers, and some dissertations; in the United States, these are indexed in sources such as the *Monthly Catalog of US Government Publications, Index to US Government Periodicals*, a whole system of legal reports (e.g. the *Pacific Reporter*, the *Supreme Court Reports*), periodical indexes (e.g. *Readers' Guide to Periodical Literature, Newsbank*), and dissertation abstracts (indexed by University Microfilms International).

A third stratum for written published texts could thus be these 'formats' represented by the various cataloguing and indexing systems. Together these indexes provide an itemized listing of published writing, and they could therefore be used as an adequate sampling frame. With a large enough sample (see following section), such a sampling frame would help achieve 'representativeness' of the various kinds of published writing. However, we know on theoretical grounds that there are several important substrata within published writing (e.g. purposes and different subject areas), and it is thus better to additionally specify these in the corpus design. This approach is more conservative, in that it insures representativeness in the desired proportions for each of these text categories, and at the same time it enables smaller sample sizes (since random techniques require larger samples than stratified techniques).

Setting and format are parallel second-level strata: format is important for the sampling of published writing; setting can be used in a similar way to provide sampling frames for unpublished writing, speech, and scripted speech. Three types of setting are distinguished here: institutional, other public, and private-personal. These settings are less adequate as sampling frames than

publication catalogues—they do not provide well defined boundaries for unpublished writing or speech, and they do not provide an exhaustive listing of texts within these categories. The problem is that no direct sampling frame exists for unpublished writing or speech. Setting, however, can be used indirectly to sample these target populations, by using three separate subcategories: institutions (offices, factories, businesses, schools, churches, hospitals, etc.), private settings (homes), and other public settings (shopping areas, recreation centres, etc.). (For scripted speech, the category of other public settings would include speech on various public media, such as news broadcasts, scripted speeches, and scripted dialogue on television sitcoms and dramas.) Operational sampling frames for each of these settings might be defined from various government and official records (e.g. census records, tax returns, or other registrations). The goal of such sampling frames would be to provide an itemized listing of the members within each setting type, so that a random sample of institutions, homes, and other public places could be selected. (These three settings could be further stratified with respect to the various types of institution, types of home, etc.)

To this point, then, I have proposed the sampling frames shown in Table 2.

Table 2 Outline of sampling frames.

Writing (published). Books/periodicals/etc. (based on available indexes)
Writing (unpublished). Institutional/public/private
Speech. Institutional/public/private
Scripted speech. Institutional/public media/other

Before proceeding, it is necessary to distinguish between two types of sampling strata. The first, as above, actually defines a sampling frame, specifying the boundaries of the operationalized population and providing an itemized listing of members. The second, as in the remaining parameters of Table 1, identifies categories that must be represented in a corpus but do not provide well-defined sampling frames. For example, Addressee plurality (no.4a: /unenumerated/plural/individual/self) provides no listing of the texts with these four types of addressee; rather, it simply specifies that texts should be collected until these four categories are adequately represented.

Further, the remaining parameters in Table 1 are not equally relevant for all the major sampling frames listed in Table 2. Consider, for example, the parameters listed under 'Addressee'. Published writing always has unenumerated addressees[5], is always written for nonpresent addressees, and is almost always non-interactive (except for published exchanges of opinion). It can require either general or specialized background knowledge (e.g. popular magazines versus academic journals) but rarely requires personal background knowledge (although this is needed for a full understanding

of memoirs, published letters, diaries, and even some novels and short stories). Unpublished writing, on the other hand, can fall into all of these addressee categories. The addressees can be unenumerated (e.g. advertisements, merchandising catalogues, government forms or announcements), plural (office circular or memo, business or technical report), individual (memo to an individual, professional or personal letter, e-mail message), or self (diary, notes, shopping list). The addressee of unpublished texts is usually absent, except in writing to oneself. Unpublished writing can be interactive (e.g. letters) or not. Finally, unpublished writing can require only general background knowledge (e.g. some advertisements), specialized knowledge (e.g. technical reports), or personal knowledge (e.g. letters and diaries).

Speech is typically directed to a plural or individual addressee, who is present. Speech addressed to self is often considered strange. Speech can be directed to unenumerated, absent addressees through mass media (e.g. a televised interview). Individual and small-group addressees can also be absent, as in the case of telephone conversations and 'conference calls'. (Individual addressees can even be non-interactive in the case of talking to an answering machine.) Private settings favour interactive addressees (either individual or small group conversations) while both interactive and non-interactive addressees can be found in institutional settings (e.g. consider the various kinds of lectures, sermons, and business presentations). General knowledge can be required in all kinds of conversation; specialized background knowledge is mostly required of addressees in institutional settings; personal knowledge is most needed in private settings.

Scripted speech is typically directed towards plural addressees (small groups in institutional settings and unenumerated audiences for mass media). Dialogue in plays and televised dramatic shows are examples of scripted speech that is directed to an individual but heard by an unenumerated audience. Addressees are typically present for scripted speech in institutional settings but are not present (physically or temporally) for scripted speech projected over mass media. Except for the lecturer who allows questions during a written speech, scripted speech is generally not interactive. Finally, scripted speech can require either general or specialized background knowledge on the part of the addressee, but it rarely requires personal background knowledge.

Addressors can vary along a number of demographic parameters (the dialect characteristics mentioned above), and decisions must be made concerning the representation of these parameters in the corpus. (Collection of texts from some addressor categories will be difficult for some sampling frames; e.g. there are relatively few published written texts by working class writers.) The second parameter here, whether the addressor is acknowledged or not, is relevant only for written texts: some written texts do not have an acknowledged personal author (e.g. advertisements, catalogues, laws and treaties, government forms, business contracts), while the more typical kinds of writing have a specific author(s).

Factuality is similar to assessments of background knowledge in that it is sometimes difficult to measure reliably, but this is an important parameter distinguishing among texts within both writing and speech. At one pole are scientific reports and lectures, which purport to be factual, and at the other are the various kinds of imaginative stories. In between these poles are a continuum of texts with different bases in fact, ranging over speculation, opinion, historical fiction, gossip, etc.

The parameter of purpose requires further research, both theoretical (as the basis for corpus design) and empirical (using the resources of large corpora). I include in Table 1 several of the purposes that should be represented in a corpus, but this is not intended as an exhaustive listing.

Similarly the parameter of topic requires further theoretical and empirical research. Library classification systems are well developed and provide adequate topic strata for published written texts. These same classifications might also serve as strata for unpublished writing, but they would need to be tested empirically. For spoken texts, especially in private settings, further research on the range of typical topics is required.

The spirit of the proposal outlined in this section is to show how basic situational parameters can be used as sampling strata to provide an important first step towards achieving representativeness. The particular parameter values used, however, must be refined, and the framework proposed here is clearly not the final word on corpus sampling strata.

3 Other sampling issues

3.1 Proportional sampling

In most stratified sample designs, the selection of observations across strata must be proportional in order to be considered representative (Williams, 1978; Henry, 1990); i.e. the number of observations in each stratum should be proportional to their numbers in the larger population. For example, a survey of citizens in North Carolina (reported in Henry, 1990, pp.61–66) used two strata, each based on a government listing of adults: households that filed 1975 income tax returns, and households that were eligible for Medicaid assistance. These two lists together accounted for an estimated 96% of the population. In the selection of observations, though, these lists were sampled proportionately—89% from the tax list and 11% from the Medicaid list—to maintain the relative proportions of these two strata in the larger population. The resulting sample can thus be claimed to represent the adult population of North Carolina. Representativeness in this case means providing the basis for accurate descriptive statistics of the entire population (e.g. average income, education, etc.).

Demographic samples are representative to the extent that they reflect the relative proportions of strata in a population. This notion of

representativeness has been developed within sociological research, where researchers aim to determine descriptive statistics that characterize the overall population (such as the population mean and population standard deviation). Any single statistic that characterizes an entire population crucially depends on a proportional sampling of strata within the population —if a strata which makes up a small proportion of the population is sampled heavily, then it will contribute an unrepresentative weight to summary descriptive statistics.

Language corpora require a different notion of representativeness, making proportional sampling inappropriate in this case. A proportional language corpus would have to be demographically organized (as discussed at the beginning of Section 3.2), because we have no a priori way to determine the relative proportions of different registers in a language. In fact, a simple demogaphically based sample of language use would be proportional by definition—the resulting corpus would contain the registers that people typically use in the actual proportions in which they are used. A corpus with this design might contain roughly 90% conversation and 3% letters and notes, with the remaining 7% divided among registers such as press reportage, popular magazines, academic prose, fiction, lectures, news broadcasts, and unpublished writing. (Very few people *ever* produce published written texts, or unpublished written and spoken texts for a large audience.) Such a corpus would permit summary descriptive statistics for the entire language represented by the corpus. These kinds of generalizations, however, are typically not of interest for linguistic research. Rather, researchers require language samples that are representative in the sense that they include the full range of linguistic variation existing in a language.

In summary, there are two main problems with proportional language corpora. First, proportional samples are representative only in that they accurately reflect the relative numerical frequencies of registers in a language—they provide no representation of relative importance that is not numerical. Registers such as books, newspapers, and news broadcasts are much more influential than their relative frequencies indicate. Secondly, proportional corpora do not provide an adequate basis for linguistic analyses, in which the range of linguistic features found in different text types is of primary interest. For example, it is not necessary to have a corpus to find out that 90% of the texts in a language are linguistically similar (because they are all conversations); rather, we want to analyse the linguistic characteristics of the other 10% of the texts, since they represent the large majority of the kinds of registers and linguistic distributions in a language.[6]

3.2 Sample size

There are many equations for determining sample size, based on the properties of the normal distribution and the sampling distribution of the mean

(or the sampling distribution of the standard deviation). One of the most important equations states that the standard error of the mean for some variable ($s_{\bar{x}}$) is equal to the standard deviation of that variable (s) divided by the square root of the sample size ($n^{1/2}$), i.e.

$$s_{\bar{x}} = s/n^{1/2}$$

The standard error of the mean indicates how far a sample mean can be from the true population mean. If the sample size is greater than 30, then the distribution of sample means has a roughly normal distribution, so that 95% of the samples taken from a population will have means that fall in the interval of plus or minus 1.96 times the standard error. The smaller this interval is, the more confidence a researcher can have that she is accurately representing the population mean. As shown by the equation for the standard error, this confidence interval depends on the natural variation of the population (estimated by the sample standard deviation) and the sample size (n). The influence of sample size in this equation is constant, regardless of the size of the standard deviation (i.e. the standard error is a function of one divided by the square-root of n). To reduce the standard error (and thus narrow the confidence interval) by half, it is necessary to increase the sample size by four times.

For example, if the sample standard deviation for the number of nouns in a text was 30, the sample mean score was 100, and the sample size was nine texts, then the standard error would be equal to 10:

$$\text{Standard error} = 30/\sqrt{(9)} = 30/3 = 10$$

This value indicates that there is a 95% probability that the true population mean for the number of nouns per text falls within the range of 80.4 to 119.6 (i.e. the sample mean of 100 ± 1.96 times the standard error of 10). To reduce this confidence interval by cutting the standard error in half, the sample size must be increased four times to 36 texts; i.e.

$$\text{Standard error} = 30/\sqrt{(36)} = 30/6 = 5$$

Similarly, if the original sample was 25 texts, then we would need to increase the sample to 100 texts in order to cut the standard error in half, i.e.

$$\text{Standard error} = 30/\sqrt{(25)} = 30/5 = 6$$

$$\text{Standard error} = 30/\sqrt{(100)} = 30/10 = 3$$

Unfortunately there are certain difficulties in using the equation for the standard error to determine the required sample size of a corpus. In particular, it is necessary to address three problems:

1 The sample size (n) depends on a prior determination of the tolerable confidence interval required for the corpus; i.e. there needs to be an apriori estimate of the amount of uncertainty that can be tolerated in typical analyses based on the corpus.
2 The equation depends on the sample standard deviation, but this is the standard deviation for some particular variable. Different variables can have different standard deviations, resulting in different estimates of the required sample size.
3 The equation must be used in a circular fashion; i.e. it is necessary to have selected a sample and computed the sample standard deviation before the equation can be used (and this is based on the assumption that the pilot sample is at least somewhat representative)—but the purpose of the equation is to determine the required sample size.

In Section 4, I consider the distribution of several linguistic features and address these three problems, making preliminary proposals regarding sample size.

3.3 A note on sampling within 'texts'

To this point I have not yet addressed the issue of how long text samples need to be. I will consider this question in more detail in Section 4, discussing the distribution of various linguistic features within texts. Here, though, I want to point out that the preference for stratified sampling applies to sampling within texts as well as across texts. Corpus compilers have typically tried to achieve better representation of texts by simply taking more words from the texts. However, these words are certainly not selected randomly (i.e. they are sequential), and the adequacy of representation thus depends on the sample length relative to the total text length. Instead it is possible to use a stratified approach for the selection of text samples from texts; i.e. especially in the case of written texts and planned spoken texts, the selection of text samples can use the typical subcomponents of texts in that register as sampling strata (e.g. chapters, sections, possibly main points in a lecture or sermon). This approach will result in better representation of the overall text, regardless of the total number of words selected from each text.

4 Distributions of linguistic features: preliminary recommendations concerning sample size

4.1 Distributions within texts: length of text samples

In this section I consider first the distribution of linguistic features within texts, as a basis for addressing the issue of optimal text length. Traditional sampling theory is less useful here than for the other aspects of corpus

design, because individual words cannot be treated as separate observations in linguistic analyses; i.e. since linguistic features commonly extend over more than one word, any random selection of words from a text would fail to represent many features and would destroy the overall structure of the text. The main issue here is thus the number of contiguous words required in text samples. The present section illustrates how this issue can be addressed through empirical investigations of the distribution of linguistic features within texts.

In Biber (1990) I approach this problem by comparing pairs of 1,000-word samples taken from single texts in the LOB and London-Lund corpora. (Text samples are 2,000 words in the LOB corpus and 5,000 words in the London–Lund corpus.) If large differences are found between the two 1,000-word samples, then we can conclude that this sample length does not adequately represent the overall linguistic characteristics of a text, and that perhaps much larger samples are required. If, on the other hand, the two 1,000-word text samples are similar linguistically, then we can conclude that relatively small samples from texts adequately represent their linguistic characteristics.

In the case of written texts (from the LOB corpus), I divided each original text in half and compared the two parts. In the case of spoken texts (from the London–Lund corpus), four 1,000-word samples were extracted from each original text, and these were then compared pairwise.

To provide a relatively broad database, ten linguistic features commonly used in variation studies were analysed. These features were chosen from different functional and grammatical classes, since each class potentially represents a different statistical distribution across text categories (see Biber, 1988). The features are: first person pronouns, third person pronouns, contractions, past tense verbs, present tense verbs, prepositions, passive constructions (combining by-passives and agentless passives), WH relative clauses, and conditional subordinate clauses. Pronouns and contractions are relatively interactive and colloquial in communicative function; nouns and prepositions are used for integrating information into texts; relative clauses and conditional subordination represent types of structural elaboration; and passives are characteristic of scientific or technical styles. These features were also chosen to represent a wide range of frequency distributions in texts, as shown in Table 3, which presents their frequencies (per 1,000 words) in a corpus of 481 spoken and written texts (taken from Biber, 1988, pp.77–78). The ten features differ considerably in both their overall average frequency of occurrence and in their range of variation. Nouns and prepositions are extremely common; present tense markers are quite common; past tense, first person pronouns, and third person pronouns are all relatively common; contractions and passives are relatively rare; and WH relative clauses and conditional subordinators are quite rare. (In addition, these features are differentially distributed across different kinds of texts; see Biber,

Table 3 Descriptive statistics for frequency scores (per 1,000 words) of ten linguistic features in a corpus of 481 texts taken from 23 spoken and written text genres.

Linguistic feature	*Mean*	*Min.*	*Max.*	*Range*
Nouns	181	84	298	214
Prepositions	111	50	209	159
Present tense	78	12	182	170
Past tense	40	0	119	119
Third person pronouns	30	0	124	124
First person pronouns	27	0	122	122
Contractions	14	0	89	89
Passives	10	0	44	44
WH relative clauses	3.5	0	26	26
Conditional subordination	2.5	0	13	13

1988, pp.246–269.) Comparison of these ten features across the 1,000-word text pairs thus represents several of the kinds of distributional patterns found in English.

The distributions of these linguistic features were analysed in 110 1,000-word text samples (i.e. fifty-five pairs of samples), taken from seven text categories: conversations, broadcasts, speeches, official documents, academic prose, general fiction, and romance fiction. These categories represent a range of communicative situations in English, differing in purpose, topic, informational focus, mode, interactiveness, formality, and production circumstances; again, the goal was to represent a broad range of frequency distributions.

Reliability coefficients were computed to assess the stability of frequency counts across the 1,000-word samples. In the case of the London-Lund corpus (the spoken texts), four 1,000-word samples were analysed from each text, and for the LOB corpus (the written texts), two 1,000-word subsamples were analysed from each text.

The reliability coefficient for each feature represents the average correlation among the frequency counts of that feature (i.e. a count for each of the subsamples). For the spoken samples, all coefficients were high. The lowest reliabilities were for passives (0.74) and conditional subordination (0.79), while all other features had reliability coefficients over 0.88. The coefficients were somewhat smaller for the written samples, in part because they are based on two instead of four subsamples. Conditional subordination in the written texts had a low reliability coefficient (0.31), while relative clauses and present tense in the written texts had moderately low reliability coefficients (0.58 and 0.61 respectively); all other features had reliability coefficients over 0.80. Overall, this analysis indicates that frequency counts for common linguistic features are relatively stable across 1,000 word samples, while frequency counts for rare features (such as conditional subordination and

WH relative clauses—see Table 3) are less stable and require longer text samples to be reliably represented.[7]

These earlier analyses can be complemented by tracking the distribution of various linguistic features across 200-word segments of texts. For example, Fig. 1 shows the distribution of prepositional phrases throughout the length of five texts from Humanities Academic Prose—the figure plots the cumulative number of prepositional phrases as measured at each 200-word interval in these texts. As can be seen from this figure, prepositional phrases are distributed linearly in these texts. That is, there are approximately the same number of prepositional phrases occurring in each 200-word segment (roughly thirty per segment in three of the texts, and twenty-five per segment in the other two texts). (The linear nature of these distributions can be confirmed by lining up a ruler next to the plot of each text.) This figure indicates that a common feature such as prepositional phrases is extremely stable in its distribution within texts (at least Humanities Academic Prose texts)—that even across 200-word segments, all segments will contain roughly the same number of prepositional phrases.

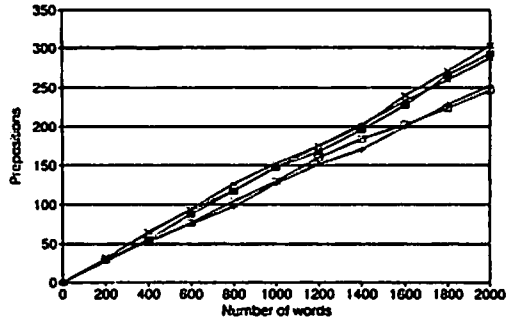


Figure 1 Distribution of prepositions in five humanities texts.

Figure 2 illustrates a curvilinear distribution, in this case the cumulative word types (i.e. the number of different words) in five Humanities texts. In general, frequency counts of a linguistic feature will be distributed linearly (although that distribution will be more or less stable within a text—see below), while frequencies of different *types* of linguistic features (lexical or grammatical) will be distributed curvilinearly; i.e. because many types are repeated across text segments, each subsequent segment contributes fewer new types than the preceding segment. In Fig. 2, the straight line marked by triangles shows the 50% boundary of word types (the score when 50% of the words in a text are different word types). In all five of these texts, at least 50% of the words are different types in the first 200 word segment (i.e. at least half of the words are not repeated), and two of the texts have more than 50% different types in the first three segments (up to 600 words). All of

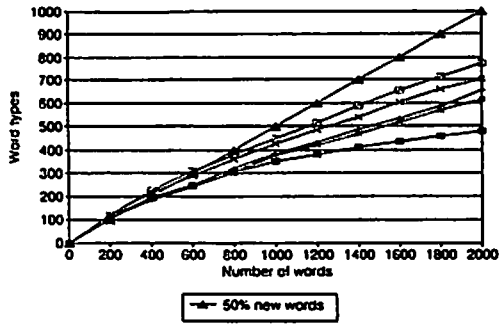


Figure 2 Distribution of word types in five humanities texts.

the texts show a gradual decay in the number of word types, however. The most diverse text drops to roughly 780 word types per 2,000 words (39%), while the least diverse text drops to roughly 480 word types per 2,000 words (only 24%). These trends would continue in longer texts, with each subsequent segment contributing fewer new types.

These two types of distributions must be treated differently. In Figs 3–9, I plot the distributions of seven linguistic features within texts representing three registers. Three of the features are cumulative frequency counts: Fig. 3 plots the frequencies of prepositional phrases, a common grammatical feature; Fig. 4 plots the frequencies of relative clauses, a relatively rare grammatical feature; and Fig. 5 plots the frequencies of noun-preposition sequences, a relatively common grammatical sequence. The other four figures plot the distributions of types in texts. Figures 6 and 7 plot the distribution of lexical types: word types (the number of different words) in Fig. 6 and hapax legomena (once-occurring words) in Fig. 7. Figures 8 and 9 plot the

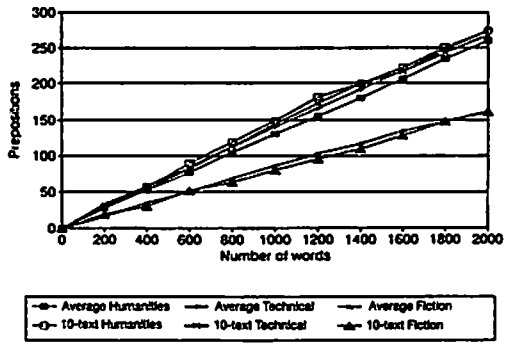


Figure 3 Distribution of prepositions in texts from three registers.

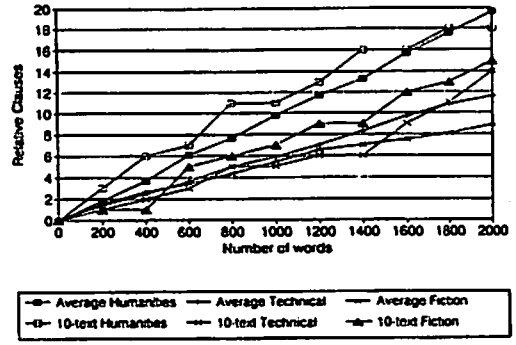


Figure 4 Distribution of relative clauses in texts from three registers.

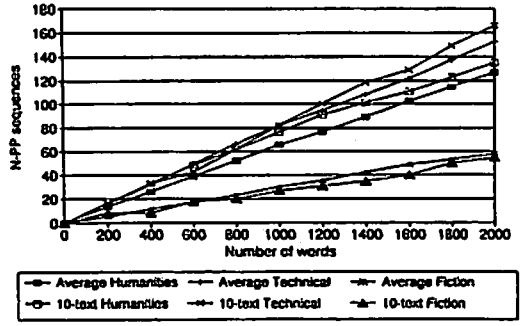


Figure 5 Distribution of N-PP sequences in texts from three registers.

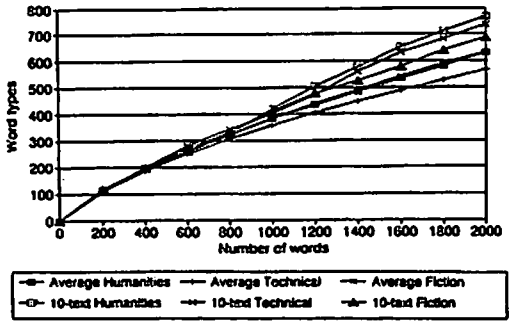


Figure 6 Distribution of word types in texts from three registers.

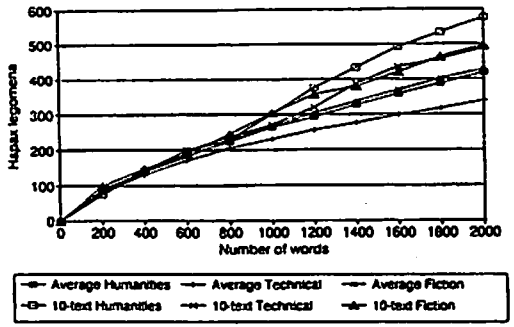


Figure 7 Distribution of *Hapax legomena* in texts from three registers.

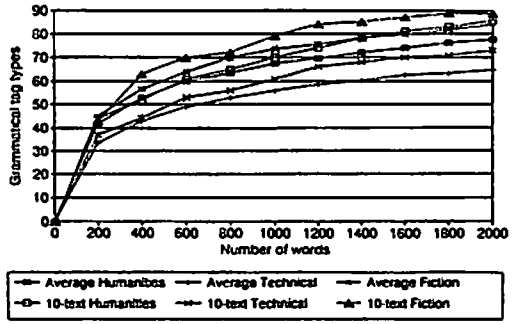


Figure 8 Distribution of grammatical tag types in texts from three registers.

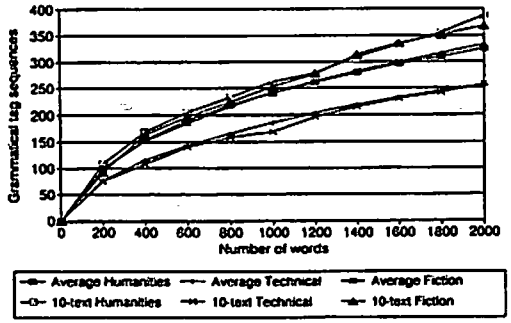


Figure 9 Distribution of grammatical tag sequences in texts from five registers.

distribution of grammatical types: different grammatical categories or 'tags' in Fig. 8, and different grammatical tag sequences in Fig. 9. The figures thus illustrate lexical and grammatical features, with rare and common overall frequencies, having linear and curvilinear distributions.

The figures can be used to address several questions. First they present the overall distributions of these features. The stable linear distribution of prepositional phrases is further confirmed by Fig. 3. In contrast, the relatively unstable distribution of relative clauses, indicated above by a relatively low reliability coefficient, is further supported by the frequent departures from linearity in Fig. 4. That is, since relative clauses are relatively rare overall, even two or three extra relatives in a 200-word segment results in an aberration. Figure 5 shows that the distribution of noun-preposition sequences is similar to that of prepositional phrases in being linear and quite stable (although less frequent overall).[8]

Figures 6–9 show different degrees of curvilinearity, with the grammatical and syntactic types showing sharper drop-offs than the lexical types. Grammatical tag types show the sharpest decay: most different grammatical categories occur in the first 200 words, with relatively few additional grammatical categories being added after 600 words.

Figures 3–9 also illustrate distributional differences across registers, although only three registers are considered here. For example, Figures 3 and 5 show fairly large differences between academic prose and fiction, with the former having much higher frequencies of prepositional phrases and noun–prepositional phrase sequences. The differences among registers are less clear-cut in Fig. 4, but humanities academic prose texts consistently have more frequent relative clauses than either technical academic prose or fiction.

Each register is plotted twice in these figures: the 'average' scores and the '10-text' scores. Average scores present the average value for ten texts from that register for the segment in question. (For example, Fig. 3 shows that humanities texts have on average 130 prepositions in the first 1,000 words of text.) In contrast, the '10-text' scores are composite scores, with each 200-word segment coming from a different text. Thus, the score for 400 words represents the cumulative totals for the first 200 words from two texts, the score for 600 words sums the first 200-word totals from three texts, etc.

In the case of stable, linear distributions, there is very little difference between the average and 10-text scores. In fact, Figs 3 and 5 show a remarkable coincidence of average and 10-text values; a single distribution is found, within a register, regardless of whether subsequent 200-word segments are taken from the same text or from different texts. Figure 4 shows greater differences for relative clauses (a relatively rare and less stable feature). Here averaging over ten texts smooths out most aberrations from linearity, while the 10-text values show considerable departures from linearity.

In contrast, there are striking differences between the average and 10-text distributions for the curvilinear features (Figs 6–9). In these cases, the

10-text scores are consistently higher than the corresponding average score from the same register. In the case of word types, the 10-text scores for all three registers are higher than the average scores of the registers. The difference is particularly striking with respect to technical academic prose—after 2,000 words of text, 10-text technical prose has the second highest word type score (approximately 740, or 37%), while on average technical prose texts have the lowest word type score (approximately 570, or 28%). This shows that there is a high degree of lexical repetition within technical prose texts, but there is a high degree of lexical diversity across technical texts. The distribution of *hapax legomena*, shown in Fig. 7, parallels that of word types—again, all three 10-text scores are higher than the average scores; 10-text humanities prose shows the highest score; and the average technical prose score is by far the lowest. These distributions reflect the considerable lexical diversity found across humanities texts, and the relatively little lexical diversity within individual technical texts.

There is more similarity between the 10-text and average scores with respect to the distribution of grammatical types (Figs 8 and 9), although for each register the 10-text score is higher than the corresponding average score. Interestingly, these figures show that technical prose has the least grammatical diversity as well as the least lexical diversity.

In summary, the analyses presented in this section indicate the following:

1 Common linear linguistic features are distributed in a quite stable fashion within texts and can thus be reliably represented by relatively short text segments.
2 Rare linguistic features show much more distributional variation within texts and thus require longer text samples for reliable representation.
3 Features distributed in a curvilinear fashion, i.e. different feature types, are relatively stable across subsequent text segments, but occurrences of new types decrease throughout the course of a text. The frequency of new types is consistently higher in cross-text samples than in single-text samples. These patterns were shown to hold for relatively short text segments (2,000 words total) and for cross-text samples taken from a single register; the patterns should be even stronger for longer text segments and for cross-register samples. These findings support the preference for stratified sampling—more diversity among the texts included in a corpus will translate into a broader representation of linguistic feature types.

With regard to the issue of text length, the thrust of the present section is simply that text samples should be long enough to reliably represent the distributions of linguistic features. For linearly distributed features, the required length depends on the overall stability of the feature. For curvilinear features, an arbitrary cut-off must be specified marking an 'adequate' representation, e.g. when subsequent text segments contribute less than 10% additional new

types. Given a finite effort invested in developing a corpus, broader linguistic representation can be achieved by focusing on diversity across texts and text types rather than by focusing on longer samples from within texts.

Specific proposals on text length require further investigations of the kind presented here, focusing especially on the distributions of less stable features, to determine the text length required for stability, and on the distributions of other kinds of features (e.g. discourse and information-packaging features).

4.2 Distributions across texts: number of texts

A second major statistical issue in building a text corpus concerns the sampling of texts: how linguistic features are distributed across texts and across registers, and how many texts must be collected for the total corpus and for each register to represent those distributions?

4.2.1 Previous research on linguistic variation within and across registers

Although registers are defined by reference to situational characteristics, they can be analysed linguistically, and there are important linguistic differences among them; at the same time, some registers also have relatively large ranges of linguistic variation internally (see Biber, 1988, Chapters 7 and 8). For this reason, the linguistic characterization of a register should include both its central tendency and its range of variation. In fact, some registers are similar in their central tendencies but differ markedly in their ranges of variation (e.g. science fiction versus general fiction, and official documents versus academic prose, where the first register of each pair has a more restricted range of variation). In Biber (1988, pp.170–198), I describe the linguistic variation within registers, including the linguistic relations among various subregisters.

The number of texts required in a corpus to represent particular registers relates directly to the extent of internal variation. In Biber (1990), I analyse the stability of feature counts across texts from a register by comparing the mean frequency counts for 10-text subsamples taken from particular registers. Five registers were analysed: conversations, public speeches, press reportage, academic prose, and general fiction. Three 10-text samples were extracted from each of these registers, and the mean frequency counts of six linguistic features were compared across the samples (first person pronouns, third person pronouns, past tense, nouns, prepositions, passives). The reliability analysis of these mean frequencies across the three 10-text samples showed an extremely high degree of stability for all six linguistic features (all coefficients greater than 0.95). These coefficients show that the mean scores of the 10-text samples are very highly correlated; that is, the central linguistic tendencies of these registers with respect to these linguistic features

are quite stable, even as measured by 10-text samples. However, there are two important issues not considered by this analysis. First, the six linguistic features considered were all relatively common; rare features such as WH relative clauses or conditional subordination might show much lower reliabilities. Secondly, this analysis addressed how many texts were needed to reliably represent mean scores, but did not address the representation of linguistic diversity in registers.[9]

4.2.2 Total linguistic variation in a corpus: total sample size for a corpus

In Section 3, I discussed how the required sample size is related to the standard error ($s_{\bar{x}}$) *by the equation:*

$$s_{\bar{x}} = s/n^{1/2} \tag{4.1}$$

The actual computation of sample size depends on a specification of the tolerable error (*te*):

$$te = t * s_{\bar{x}} \tag{4.2}$$

Equation 4.2 states that the tolerable error is equal to the standard error times the *t*-value. Given a sample size greater than thirty (which permits the assumption of a normal distribution), a researcher can know with 95% confidence that the mean score of a sample will fall in the interval of the true population mean plus-or-minus the tolerable error.

Equation 4.2 can be manipulated to provide a second equation for computing the standard error, i.e. $s_{\bar{x}} = te/t$. If the ratio *te/t* is substituted for $s_{\bar{x}}$ in Equation 4.1, and the equation is then solved for *n*, we get a direct computation of the required sample size for a corpus:

$$n = s^2/(te/t)^2 \tag{4.3}$$

where *n* is the computed sample size, *s* is the estimated standard deviation for the population, *te* is the tolerable error (equal to 1/2 of the desired confidence interval), and *t* is the *t*-value for the desired probability level.

I note in Section 3 that there are problems in the application of Equation 4.3. In one sense, the equation simply shifts the burden of responsibility, from estimating the unknown quantity for required sample size to estimating the unknown quantities for the tolerable error and population standard deviation; i.e. in order to use the equation, there needs to be a prior estimate of the tolerable error or confidence interval permitted in the corpus and a prior estimate of the standard deviation of variables in the population as a whole.

The tolerable error depends on the precision required of population estimates based on the corpus sample. For example, say that we want to

know how many nouns on average occur in conversation texts. The confidence interval is the window within which we can be 95% certain that the true population mean falls. For example, if the sample mean for nouns in conversations was 120, and we needed to estimate the true population mean of nouns with a precision of ± 2, then the confidence interval would be 4, extending from 118 to 122. The tolerable error is simply one side (or one-half) of the confidence interval. The problem here is that it is difficult to provide an a priori estimate of the required precision of the analyses that will be based on a corpus.

Similar problems arise with the estimation of standard deviations. In this case, it is not possible to estimate the standard deviation of a variable in a corpus without already having a representative sample of texts. Here, as in many aspects of corpus design, work must proceed in a circular fashion, with empirical investigations based on pilot corpora informing the design process. The problem for initial corpus design, however, is to provide an initial estimate of standard deviation.

A final problem is that standard deviations must be estimated for particular variables, but in the case of corpus linguistics, there are numerous linguistic variables of interest. Choosing different variables, with different standard deviations, will result in different estimates of required sample size.

In the present section, I use the analyses in Biber (1988, pp.77–78, 246–269) to address the first two of these problems. That study is based on a relatively large and wide-ranging corpus of English texts: 481 texts taken from twenty-three spoken and written registers. Statistical analyses of this corpus can thus be used to provide initial estimates for both the tolerable error and the population standard deviation.

In the design of a text corpus, tolerable error cannot be stated in absolute terms because the magnitude of frequency counts varies considerably across features (as was shown in Section 3). For example, a tolerable error of ± 5 might work well for common features such as nouns, which have an overall mean of 180.5 per 1,000 words in the pilot corpus, but it would be unacceptable for rare features such as conditional subordinate clauses, which have an overall mean of only 2.5 in the corpus (so that a tolerable error of 5 would translate into a confidence interval of –2.5 to 7.5, and a text could have three times the average number of conditional clauses and still be within the confidence interval). Instead I propose here computing a separate estimate of the tolerable error for each linguistic feature, based on the magnitude of the mean score for the feature; for illustration, I will specify the tolerable error as ± 5% of the mean score (for a total confidence interval of 10% of the mean score). Table 4 presents the mean score and standard deviation of seven linguistic features in the pilot corpus, together with the computed tolerable error for each feature. It can be seen that the tolerable error ranges from 9.03 for nouns (which have a mean of 180.5) to 0.13 for conditional clauses (which have a mean of only 2.5).

Table 4 Estimates of required sample sizes (number of texts) for the total corpus.

	Mean score in pilot corpus	*Standard deviation in pilot corpus*	*Tolerable error*	*Required* N
Nouns	180.5	35.6	9.03	59.8
Prepositions	110.5	25.4	5.53	81.2
Present tense	77.7	34.3	3.89	299.4
Past tense	40.1	30.4	2.01	883.1
Passives	9.6	6.6	0.48	726.3
WH relative clauses	3.5	1.9	0.18	452.8
Conditional clauses	2.5	2.2	0.13	1,190.0

Given the tolerable errors and estimated standard deviations listed in Table 4, required sample size (i.e. the total number of texts to be included in the corpus) can be computed directly using Equation 4.3. Table 4 shows very large differences in required sample size across these linguistic features. These differences are a function of the size of the standard deviation relative to the mean for a particular feature. If the standard deviation is many times smaller than the mean, as in the case of common features such as nouns and prepositions, the required sample size is quite small. If, on the other hand, the standard deviation approaches the mean in magnitude, as in the case of rare features such as WH relative clauses and conditional clauses, the required sample size becomes quite large. Past tense markers are interesting in that they are relatively common (mean of 40.1) yet have a relatively large standard deviation (30.4) and thus require a relatively large sample of texts for representation (883). Overall the most conservative approach in designing a corpus would be to use the most widely varying feature (proportional to its mean—in this case conditional clauses) to set the total sample size.

4.2.3 Linguistic variation within registers: number of texts needed to represent registers

The remaining issue concerns the required sample size for each register. Although most books on sample design simply recommend proportional sampling for stratified designs (see Section 3), a few books discuss the need for non-proportional stratified sampling in certain instances; these books differ, however, on the method for determining the recommended sample sizes for subgroups. For example, Sudman (1976, pp.110–111) states that non-proportional stratified sampling should be used when the subgroups themselves are of primary interest (as in the case of a text corpus), and that the sample sizes of the subgroups should be equal in that case (to minimize the standard error of the difference). This procedure is appropriate when the variances of the subgroups are roughly the same. In contrast, Kalton (1983,

pp.24–25) recommends using the subgroup standard deviations to determine their relative sample sizes. This procedure is more appropriate for corpus design, since the standard deviations of linguistic features vary considerably from one register to the next.

Although I do not make specific recommendations for register sample size here, I illustrate this approach in Table 5, considering the relative

Table 5 Relative variation within selected registers.

Conversations. Average normalized deviation = 0.37

	Mean score in pilot corpus	*Standard deviation in pilot corpus*	*Ratio of standard deviation/mean (normalized deviation)*
Nouns	137.4	15.6	0.11
Prepositions	85.0	12.4	0.15
Present tense	128.4	22.2	0.17
Past tense	37.4	17.3	0.46
Passives	4.2	2.1	0.50
WH Relative clauses	1.4	0.9	0.64
Conditional clauses	3.9	2.1	0.54

General fiction. Average normalized deviation = 0.39

	Mean score in pilot corpus	*Standard deviation in pilot corpus*	*Ratio of standard deviation/mean (normalized deviation)*
Nouns	160.7	25.7	0.16
Prepositions	92.8	15.8	0.17
Present tense	53.4	18.8	0.35
Past tense	85.6	15.7	0.18
Passives	5.7	3.2	0.56
WH Relative clauses	1.9	1.1	0.58
Conditional clauses	2.6	1.9	0.73

Academic prose. Average normalized deviation = 0.49

	Mean score in pilot corpus	*Standard deviation in pilot corpus*	*Ratio of standard deviation/mean (normalized deviation)*
Nouns	188.1	24.0	0.13
Prepositions	139.5	16.7	0.12
Present tense	63.7	23.1	0.36
Past tense	21.9	21.1	0.96
Passives	17.0	7.4	0.44
WH Relative clauses	4.6	1.9	0.41
Conditional clauses	2.1	2.1	1.00

variances of seven linguistic features (the same as in Table 4) across three registers: conversations, general fiction, and academic prose. As above, the data are taken from Biber (1988, pp.246–269).

Table 5 presents the mean score, standard deviation, and the ratio of standard deviation to mean score, for these seven linguistic features in the three registers. The ratio represents the normalized variance of each of these features within each register—the extent of internal variation relative to the magnitude of the mean score. The raw standard deviation is not appropriate here (similar to Table 4) because the mean scores of these features vary to such a large extent.

Table 5 shows that the normalized standard deviation varies considerably across features within a register. For example, within conversations the counts for nouns, prepositions, and present tense all show relatively small normalized variances, while passives, WH relative clauses, and conditional clauses all show normalized variances at or above 50%. As shown earlier, features with lower overall frequencies tend to have considerably higher normalized variances.

There are also large differences across the registers. For example, past tense has a normalized variance of 46% in conversations and only 18% in general fiction, but it shows a normalized variance of 96% in academic prose. Conditional subordination also shows large differences across these three registers: it has a normalized variance of 54% in conversations, 73% in general fiction, and 100% in academic prose.

In order to determine the sample size for each register, it is necessary to compute a single measure of the variance within each register. This measure is then used to allot a proportionally larger sample to registers with greater variances. (This should not be confused with a proportional representation of the registers.) A certain minimum number of texts should be allotted for each register (e.g. at least twenty texts per register), and then the remaining texts in the corpus can be divided proportionally depending on the relative variance within registers.

To illustrate, consider Table 5 again. This table lists an average normalized deviation for each register, which represents an overall deviation score computed by averaging the normalized standard deviations of the seven linguistic features. Conversations and general fiction both have relatively similar overall deviations (37% and 39% correspondingly) whle academic prose has a somewhat higher overall deviation (49%). To follow through with this example, assume that there were to be a total of 200 texts in a corpus, taken from these three registers. Each register would be allotted a minimum of twenty texts, leaving 140 texts to be divided proportionally among the three registers. To determine the relative sample size of the registers, one would solve the following equation based upon their relative overall deviations:

$$0.37x + 0.39x + 0.49x = 140$$
$$1.25x = 140$$
$$x = 112$$

and thus the sample sizes would be:

Conversation 0.37 * 112 = 41
General fiction 0.39 * 112 = 44
Academic prose 0.49 * 112 = 55
Total allocated texts = (41 + 20 for conversation) + (44 + 20 for general fiction) + (55 + 20 for academic prose) = 200 texts

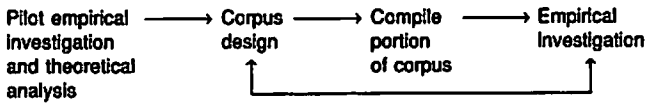
To compute the actual values for register sample sizes, it is necessary to analyse the full range of linguistic features in all registers, computing a single average deviation score for each register. This could be done by averaging across the normalized variances of all linguistic features, as illustrated here. An alternative approach would be to use the normalized variances of the linguistic dimensions identified in Biber (1988). This latter approach would have a more solid theoretical foundation, in that the dimensions represent basic parameters of variation among registers, each based on an important co-occurrence pattern among linguistic features. In contrast, the approach illustrated in this section depends on the pooled influence of linguistic features in isolation, and thus relatively aberrant distributions can have a relatively strong influence on the final outcome. In addition, use of the dimensions enables consideration of the distributions with respect to particular functional parameters, so that some dimensions can be given more weight than others. In contrast, there is no motivated way for distinguishing among the range of individual features on functional grounds.

It is beyond the scope of this paper to illustrate the use of dimension scores for the linguistic characterization of registers (since I would first need to explain the theoretical and methodological bases of the dimensions). The same basic approach as illustrated in this section would be used, however. The major difference involves the analysis of deviation along basic dimensions of linguistic variation rather than with respect to numerous linguistic features in isolation.

5 Conclusion: beginning

I have tried to develop here a set of principles for achieving 'representativeness' in corpus design. I have offered specific recommendations regarding some aspects of corpus design, and illustrations elsewhere (regarding issues for which final recommendations could not be developed in a paper of this

scope). The bottom-line in corpus design, however, is that the parameters of a fully representative corpus cannot be determined at the outset. Rather, corpus work proceeds in a cyclical fashion that can be schematically represented as follows:



Theoretical research should always precede the initial corpus design and actual compilation of texts. Certain kinds of research can be well advanced prior to any empirical investigations: identifying the situational parameters that distinguish among texts in a speech community, and identifying the range of important linguistic features that will be analysed in the corpus. Other design issues, though, depend on a pilot corpus of texts for preliminary investigations. Present-day researchers on English language corpora are extremely fortunate in that they have corpora such as the Brown, LOB, and London–Lund corpus for pilot investigations, providing a solid empirical foundation for initial corpus design. The compilers of those corpora had no such pilot corpus to guide their designs. Similar situations exist for current projects designing corpora to represent non-western languages. For example, a recently completed corpus of Somali required extensive fieldwork to guide the initial design (see Biber and Hared, 1992). Thus the initial design of a corpus will be more or less advanced depending on the availability of previous research and corpora.

Regardless of the initial design, the compilation of a representative corpus should proceed in a cyclical fashion: a pilot corpus should be compiled first, representing a relatively broad range of variation but also representing depth in some registers and texts. Grammatical tagging should be carried out on these texts, as a basis for empirical investigations. Then empirical research should be carried out on this pilot corpus to confirm or modify the various design parameters. Parts of this cycle could be carried out in an almost continuous fashion, with new texts being analysed as they become available, but there should also be discrete stages of extensive empirical investigation and revision of the corpus design.

Finally, it should be noted that various multivariate techniques could be profitably used for these empirical investigations. In this paper, I have restricted myself to univariate techniques, and to simple descriptive statistics. Other research, though, suggests the usefulness of two multivariate techniques for the analysis of linguistic variation in computerized corpora: factor analysis and cluster analysis. Factor analysis can be used in either an exploratory fashion (e.g. Biber, 1988) or for theory-based 'confirmatory' analyses (e.g. Biber, 1992). Both of these would be appropriate for corpus

design work, especially for the analysis of the range and types of variation within a corpus and within registers. Such analyses would indicate whether the different parameters of variation were equally well represented and would provide a basis for decisions on sample size. Cluster analysis has been used to identify 'text types' in English—text categories defined in strictly linguistic terms (Biber, 1989). Text types cannot be identified on a priori grounds; rather they represent the groupings of texts in a corpus that are similar in their linguistic characterizations, regardless of their register categories. Ideally a corpus would represent both the range of registers and the range of text types in a language, and thus research on variation within and across both kinds of text categories is needed.[10]

In sum, the design of a representative corpus is not truly finalized until the corpus is completed, and analyses of the parameters of variation are required throughout the process of corpus development in order to fine-tune the representativeness of the resulting collection of texts.

Notes

1 Further, in the case of language corpora, proportional representation of texts is usually not desirable (see Section 3); rather, representation of the range of text types is required as a basis for linguistic analyses, making a stratified sample even more essential.

2 Actually, very little work has been carried out on dialect variation from a text-based perspective. Rather, dialect studies have tended to concentrate on phonological variation, downplaying the importance of grammatical and discourse features.

3 Other demographic factors characterize individual speakers and writers rather than groups of users; these include relatively stable characteristics, such as personality, interests, and beliefs, and temporary characteristics, such as mood and emotional state. These factors are probably not important for corpus design, unless an intended use of the corpus is investigation of personal differences.

4 This parameter would not be important for many nonwestern societies, or for certain kinds of corpora representing different historical periods; quite different sampling strategies would be required in these cases.

5 Published collections of letters and published diaries are special cases—these originally have individual addressees, but they are usually written with the hope of eventual publication and thus with an unenumerated audience in mind.

6 A proportional corpus would be useful for assessments that a word or syntactic construction is 'common' or 'rare' (as in lexicographic applications). Unfortunately, most rare words would not appear at all in a proportional (i.e. primarily conversational) corpus, making the database ill-suited for lexicographic research.

7 In Biber (1990) I also assess the representativeness of 1,000-word text samples by computing difference scores for pairs of samples from each text. This analysis confirms the general picture given by the reliability coefficients while providing further details of the distribution of particular features in particular registers.

8 These are primarily prepositional phrases functioning as noun modifiers, as opposed to prepositional phrases with adverbial functions.

9 Actually this latter question was addressed by computing difference scores, for the mean, standard deviation, and range, across the 10-text samples.

10 For example, one of the most marked text types identified in Biber (1989) consists of texts in which the addresser is producing an on-line reportage of events in progress. Linguistically, this text type is marked in being extremely situated in reference (many time and place adverbials and a present time orientation). Unfortunately, there are only seven such texts in the combined London–Lund and LOB corpora, indicating that this text type is under-represented and needs to be targeted in future corpus development.

Acknowledgements

I would like to thank Edward Finegan for his many helpful comments on an earlier draft of this paper. A modified version of this paper was distributed for the Pisa Workshop on Textual Corpora, held at the University of Pisa (January 1992), and discussions with several of the workshop participants were also helpful in revising the paper.

References

Biber, D. (1986). Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings. *Language*, 62: 384–414.

—— (1988). *Variation across Speech and Writing*. Cambridge University Press: Cambridge.

—— (1989). A Typology of English Texts, *Linguistics*, 27: 3–43.

—— (1990). Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. *Literary and Linguistic Computing*, 5: 257–269.

—— (1992). On the Complexity of Discourse Complexity: A Multidimensional Analysis. *Discourse Processes*. 15: 133–163.

—— (1993a). An Analytical Framework for Register Studies. In D. Biber and E. Finegan (eds), *Sociolinguistic Perspectives on Register*. Oxford University Press, New York. In press.

—— (1993b). Register Variation and Corpus Design, *Computational Linguistics*. In press.

—— and Hared, M. (1992). Dimensions of Register Variation in Somali, *Language Variation and Change*, 4: 41–75.

Brown, P. and Fraser, C. (1979). Speech as a Marker of Situation. In K. R. Scherer and H. Giles (eds), *Social Markers in Speech*. Cambridge University Press, Cambridge, pp.33–62.

Duranti, A. (1985). Sociocultural Dimensions of Discourse. In T. van Dijk (ed.), *Handbook of Discourse Analysis*, Vol. 1. Academic Press: New York: pp.193–230.

Francis, W. N. and Kucera, H. (1964/1979). *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Department of Linguistics, Brown University.

Halliday, M. A. K. and Hasan, R. (1989). *Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective*. Oxford University Press, Oxford.

Henry, G. T. (1990). *Practical Sampling*. Sage, Newbury Park, CA.

Hymes, D. H. (1974). *Foundations in Sociolinguistics*. University of Pennsylvania Press, Philadelphia.

Johansson, S., Leech, G. N. and Goodluck, H. (1978). Manual of information to accompany the Lancaster–Oslo/Bergen Corpus of British English, for use with digital computers. Department of English, University of Oslo.

Kalton, G. (1983). *Introduction to survey sampling*. Sage, Newbury Park, CA.

Sudman, S. (1976). *Applied Sampling*. Academic Press, New York.

Svartvik, J. and Quirk, R. (eds) (1980). *A Corpus of English Conversation*. C. W. K. Gleerup, Lund.

Williams, B. (1978). *A Sampler on Sampling*. John Wiley and Sons, New York.