II. Corpus compilation and corpus types

9. Collection strategies and design decisions

1. Introduction

Corpora range in type from general, reference corpora designed to investigate a given language as a whole, to specialised corpora designed to answer more specific research questions (cf. article 3). They can be carefully planned and have a long 'shelf-life', or they can be 'disposable' (Bernardini/Baroni 2004), quickly constructed for a specific purpose and as rapidly discarded. This article deals with some of the issues involved in planning and compiling a corpus. As will be seen in this article, this is an area prone to paradox, where even the apparently simplest decisions can have extensive ramifications.

Any corpus, unless it is unusually specific in content, may be perceived as a collection of sub-corpora, each one of which is relatively homogenous. The sub-corpora are determined by a template of variables that creates a number of cells, each of which constitutes a sub-corpus. For example, a researcher may wish to investigate the written language of Economics and of Social Science, and may wish to include academic articles, university textbooks and popular articles in the corpus. These design criteria may be expressed as a grid, cf. Table 9.1.

Tab. 9.1: A grid of corpus design criteria

	Economics	Social Sciences
Academic articles	1	2
University textbooks	3	4
Popular articles	5	6
Student essays	7	8

The corpus compiler collects texts to fit into each of the eight cells and may wish to specify, for example, the relative size of each sub-corpus in terms of texts, or of tokens.

If the corpus is being built for a specific purpose, as in this example, the variables will be determined by the parameters of the study – in this case the distinction between two academic disciplines and the genres that constitute them. If not, the criteria may be

drawn from theories of language variation. Aston and Burnard (1998, 23; 29−33), for example, refer to systemic linguistic concepts of *field*, *tenor* and *mode* variables; these are (just) recognisable in the design criteria of the British National Corpus (BNC) which include Domain (i. e. the subject-matter of written texts), Interaction type (for the spoken texts), and Medium (e. g. book, periodical, written-to-be-spoken). On the other hand, lists of variables are more often recognisable as 'common-sense' distinctions rather than as the outcomes of any one given theory. For example Nelson/Wallis/Aarts (2002, 307−308) describe the British component of the International Corpus of English (ICE-GB) as being subdivided into:

Spoken vs. Written
(In spoken) Dialogue vs. Monologue
(In dialogue) Private vs. Public
(In private) Face-to-face conversation vs. Telephone

and so on. Other variables often used in sociolinguistics, such as age, sex, region and social class are also used as design criteria.

The purpose of identifying and filling each cell in the template may be to allow comparison between them. Biber (1988, 3), for example, concentrates on this aspect. Alternatively, it may simply be to ensure that the widest possible range of texts is included. Our hypothetical researcher into Economics and Social Sciences, for example, may wish to compare student essays in each discipline, or to compare academic articles with popular articles, in both disciplines. Equally possibly, however, he or she may wish to compile a lexicon for students of these subjects, and may include a number of different genres only in order to ensure that no crucial items of vocabulary are missed.

Many corpora are, of course, much more complex than the example above, and the hypothetical grid may quickly become multi-dimensional. (Our hypothetical researcher may choose to distinguish between texts published in Britain and in the US, for example, and/or between textbooks aimed at different kinds of students.) Much of the debate over corpus design focuses on how the parameters of the grid are to be determined, and what the relationship between the cells should be. This article will consider such questions, and is organised as follows: section 2 makes the point that a corpus cannot be judged except in the context of its purpose; sections 3, 4 and 5 identify some practical constraints on corpus design, and on the construction of written and spoken corpora; sections 6, 7 and 8 discuss some theoretical issues in corpus design.

2. Corpus use and corpus design

If a corpus is compiled in order to carry out research only on its own content, identifying what that content should be is straightforward. For example, Barr (2003) carries out a stylometrics study comparing texts from the New Testament, using a corpus comprising those texts. In many cases, however, it is clearly impossible to put all the texts that are the object of research into a corpus. Even a fairly restricted topic of study − lectures delivered in American universities during the year 2004, for example − involves too many texts for a corpus of all the relevant texts to be compiled. Instead a sub-set of the possible candidate texts is selected, and in that selection lies corpus design.

It is a truism that there is no such thing as a 'good' or a 'bad' corpus, because how a corpus is designed depends on what kind of corpus it is and how it is going to be used.

Articles 10 to 17 give detailed information about types of corpora. Here, I shall simply give some examples to show how the design of a corpus depends on its purpose.

A corpus intended to facilitate research into a single register, such as university lectures, will contain texts from the range represented by that register. It is likely that researchers will wish to compare lectures in different disciplines and/or introductory lectures with more advanced ones. The corpus designer may decide to select as many disciplines as resources will allow, and to include lectures from all stages of the students' career in each discipline. Even if some disciplines are more 'lecture based' than others, so that in the university as a whole more Law lectures, say, are given each year than Geography ones, the corpus designer may choose to include roughly equal numbers of each to make comparison easier. In other words, strict sampling techniques may give way to the need for comparability (cf. sections 6 and 7).

In an example such as this one, the texts are selected on external criteria. For example, a text may be included because it is a transcript of a lecture given in the Law department to first year students, not because it covers a particular topic or because it contains instances of a particular word or phrase. The design of most corpora is based on such external criteria, that is, using situational determinants rather than linguistic characteristics, as the parameters of composition (Biber 1988, 68). However, where the purpose of the research project is to discover how a particular cultural keyword (Stubbs 1996, 157–195) is used, or how a word has changed its meaning or function over a period of time (Teubert 2004, 138–155), a corpus consisting of texts selected precisely because they contain that item might be justified. This is particularly true if the item concerned is relatively infrequent, or if data from a specific time period is required. In these cases even a very large general corpus may be inadequate to represent the item in its required contexts. For example, to test the hypothesis that the affective meaning of the words *unilateral, unilateralism, unilaterally* is different in British and American English, and that it has changed since 2001, requires a corpus of texts containing those words, from the two countries, covering a number of specific years, and of a sufficient size to allow comparison to take place. (This example is based on Rottweiler 2006.) The texts therefore have to be selected using both internal criteria (they contain at least one of the target words at least once) and external criteria (there are an approximately equal number originating in Britain or America, and in each of the years 1999–2003).

The contents of a corpus designed for research purposes, whether general or specialised, need to be carefully considered. On the other hand, a language teacher may wish to compile a small corpus for his or her students to use in checking how particular words and phrases in the target language are typically used. The students may not need a corpus that is a balanced representation of the language as a whole, but a ready reference that can be cross-checked against books and the teacher's intuition where necessary. A corpus of newspaper texts on CD-ROM, or texts downloaded from the Internet, will be a sufficient source of information about how the most frequent words and phrases in the language are used. Design of the corpus will depend more on what is freely available in an easily-converted format than on other criteria.

3. Practical constraints on corpus design

All corpora are a compromise between what is desirable, that is, what the corpus designer has planned, and what is possible. There are many practical constraints on corpus

building, of which the most important are: software limitations, copyright and ethical issues, and text availability. Each of these will be dealt with briefly.

Useful corpus size may be limited by the search software that is to be used. Readily available software packages, such as WordSmith Tools (Scott 2004), work on raw text, and can deal with a corpus of tens of millions of tokens in size. Larger corpora often work with software that demands that the tokens are converted into digits (each type being represented by a unique sequence of digits). This enables the search software to work more quickly. Even so, complex operations on corpora of many hundreds of millions of words can take some time to complete.

Storing electronic versions of published texts is illegal in most countries unless copyright permission has been given. Such permission is often difficult to obtain, and even when given may restrict the availability of the corpus (Meyer 2002, 62). The design of many corpora is determined to a large extent by the availability or otherwise of copyright permissions. Corpora consisting of unpublished material, such as student essays or transcripts of conversations, do not run into copyright problems as such, but ethical considerations must be taken into account. Informed consent for the material to be used must be obtained, and this has to include permission for the corpus to be made publicly available, if this is what is intended. Ensuring anonymity for participants in spoken interactions is possible (for example, names can be changed or deleted in transcripts: Meyer 2002, 75), but not if the transcripts are linked to sound files.

The third practical issue to be taken into account is the availability of texts, and their availability in a usable form. Historical corpora, in particular, are constrained by the limitations on what texts from earlier times are currently available. It is often pointed out, for example, that obtaining spoken texts from an era before the invention of tape recorders is impossible, and corpus designers hoping to include such material must depend on transcripts of situations such as court proceedings, or on fictional representations of speech (Biber and Finegan 2001, 69; Meyer 2002, 37). In either case, the accuracy of the data as a representation of actual speech is always questionable. Written texts are easier to obtain, but in some cases they may be available only in paper form rather than electronically. Unless very large resources are available for scanning and keying in (cf. section 4), a corpus designer may choose to avoid texts that are not available electronically. Thus, once again, ideal corpus design may take second place to availability (see article 14).

Practical constraints operate also in cases where a corpus is designed to be non-finite and where it will be added to over time in order to track changes in a language (i. e. a 'monitor corpus' as described by Sinclair 1991, 24–26 and Teubert 2003, 12). It is unlikely that the resources necessary for compiling a carefully balanced and varied corpus will be available in perpetuity, and monitor corpora may need to make use of texts that are available cheaply and easily, relying on internet and/or journalistic texts.

4. Collecting written texts

There are three methods of acquiring written texts in a form that can be used to create a corpus. In increasing level of technological sophistication, and ease, they are: keying-in, scanning, and obtaining texts electronically.

Keying texts in by hand is obviously very time-consuming and is generally avoided unless the texts concerned are unavailable in any other way, as may be the case, for example, with older manuscripts, or handwritten letters, or learner essays. During the keying-in process, decisions may have to be taken, for example, whether to normalise unconventional spelling.

Scanning was the usual method of building corpora in the 1980s (Renouf 1987, 5) and is still used where print quality is sufficient, and where a text cannot be obtained in electronic form. With the larger corpora that are expected today, however, obtaining text in electronic form, either from a publisher or from the internet, is the optimum way of building a corpus of written texts.

It is no exaggeration that the availability of the internet, with its instant access to millions of downloadable texts, has transformed corpus building (see article 18). In most cases, corpus builders find the internet a more convenient source of texts than material in paper form. Bernardini/Baroni (2004), for example, describe software designed to trawl the internet and compile a 'quick and dirty' corpus of texts on topics chosen by the compiler. Such a corpus is not designed as a permanent research tool, but as a useful temporary aid for a language learner seeking to extend their familiarity with lexis in a given domain. These 'disposable' corpora are a far cry from the carefully designed and painstakingly constructed corpora of the twentieth century. Meyer (2002, 63) also approves of the relative ease of building a corpus using internet texts, but warns that such texts may not be identical to those that appear in print, so that an 'internet corpus' is an artefact of a particular kind, representing language on the internet, not written language in general.

Teubert (2001, 45–46) takes a more robust attitude. Describing a corpus of texts produced by British Euro-sceptics (those who oppose Britain's membership of the European Union) and posted on web-sites, he argues that such texts, because of their ready availability to the average web-user, are more influential upon public discourse than those that could be found in newspapers or in more esoteric publications such as Hansard (the record of proceedings in the British parliament). Compiling a corpus exclusively of such texts, therefore, is not simply a matter of convenience, but of policy.

5. Collecting spoken texts

The larger corpora become, therefore, the more corpus builders tend to rely on material that is easily available in written form. This is at odds with the procedures necessary for collecting spoken data. There are several well-known spoken corpora (see articles 11 and 30 for a description of many of them). Many are relatively specialised, focusing on the interactions of particular sections of a community. More ambitious in design is the spoken component of the British National Corpus, which attempts to include speakers of all ages, socio-economic groups, and regions in Britain, and to represent a wide range of interaction types (Leech/Rayson/Wilson 2001, 2–4). The compilers of the British spoken sub-corpus of the Bank of English took a more serendipitous approach to build as large a sub-corpus as possible (20 million words): it contains a wide variety of social situations, such as casual conversation among friends, seminars, meetings, service encounters, unscripted local radio broadcasts, and interviews conducted by researchers in

History and Sociology, but there is no attempt to control for a balance of gender, age, region or class.

Three questions face the compiler of a corpus of spoken language. One is the selection of speakers and social contexts; another is the management of data collection; the third is the choice of transcription system. Some aspects of the selection of speakers and contexts might be determined by the aim of the corpus. For example, the age range of speakers in the COLT corpus is limited by the requirement to obtain teenage language. The Santa Barbara corpus includes only casual conversation. Many learner corpora, designed to investigate the interlanguage of learners, take oral language tests, which are usually tape-recorded as a matter of course, as their component texts. These have the advantage of adding a measure of uniformity to the texts, as the learners are performing similar tasks. On the other hand, what is investigated is the language of learners under test conditions rather than the totality of their production.

Those projects, such as the BNC, that attempt to represent the spoken language of a nation, require ingenuity to overcome the inevitable difficulties. The compilers of the BNC used market-research interviewers to identify a cross-section of speakers who would be willing to record themselves over a period of time. They were therefore drawing on the expertise and recognised procedures of a profession accustomed to sampling populations. As Aston and Burnard (1998, 31) make clear, however, such sampling was successful in obtaining interactions in some social situations only, and further collection, not sampled demographically, had to be done in order to obtain examples of lectures, legal proceedings, radio broadcasts and so on. The result is a corpus of two halves, one balanced in terms of speaker age, sex and so on only, the other balanced in terms of interaction type only.

The questions of data collection and transcription are similar to those faced by any researcher into spoken interaction, but exacerbated by the need to have a relatively large number of texts. Meyer (2002, 56–61) describes some of these: the need to obtain informed consent from all speakers, the choice of recording equipment, and the problems caused by the speakers' awareness that they are being recorded. He notes that some corpus compilers have adopted techniques such as giving target speakers recording equipment and asking them to turn on the recorder whenever they wish. This avoids the difficulty of requiring very large numbers of researchers to obtain the necessary amount of data. A corpus also puts constraints upon transcription systems. Because of the need to search the corpus by entering a search word orthographically, normalised spelling is usually used, rather than the spelling representing sound used by conversation analysts. Timing features such as overlaps need to be represented through symbol rather than by the layout of the text, as the corpus will be stored purely linearly. Other than that, the amount of information encoded in the transcript depends largely on the size of the corpus. Meyer (2002, 71) notes two extremes: the Corpus of Spoken Professional English ⟨http://www.athel.com/corpdes.html⟩ which consists of ready-made transcripts produced by professional transcribers who were nonetheless not transcribing for the purposes of linguistic analysis, and the Santa Barbara Corpus, which is a faithful transcription, of the standard expected for conversation analysis, including hesitations, false starts and also information about intonation. The time and level of expertise needed to undertake such a detailed transcription means that the corpus is of necessity relatively small.

As Meyer (2002, 72) points out, a written representation of speech can be only partially informative, however accurate the transcription, and this is why many corpora of

spoken interaction are now linked to sound files. However, corpus linguistics has yet to embrace fully the issues involved in transferring its principles to interactions in other media. The mainstays of corpus research, such as concordance lines and word lists, assume a written medium. Other media, such as film, will need different methods of presenting data. Baldry/Thibault (2005), for example, report a multi-media corpus that is actually a collection of video-clips, heavily annotated so that they can be searched for specific instances of given semiotic categories. The need to devise methodologies for exploring multi-media corpora is particularly acute when sign languages are being studied, as these do not have an accepted written form and can currently be studied only through video-recordings of signed texts.

6. The corpus and the object of investigation

There are three issues which are typically taken into account when designing a corpus. These are sometimes referred to as representativeness, balance, and size. They will now be discussed in turn, in each case taking an example of a publicly-available corpus.

Representativeness is the relationship between the corpus and the body of language it is being used to represent. A corpus is usually intended to be a microcosm of a larger phenomenon, except where the corpus is the whole, as in the Barr (2003) example mentioned above. As such, although some statements can be made with absolute confidence about the corpus itself, the value of the corpus lies in being able to make somewhat more tentative statements about the body of language as a whole. Thus a corpus that is unrepresentative is very limited in usefulness.

For example, the ICLE corpus and the LOCNESS corpus (Granger 1998, 9–10; 13) both consist of expository essays written by university students in English. In the case of the ICLE corpus, the writers are learners of English, whereas the LOCNESS writers are native speakers of English. Aijmer (2002, 61) reports that the Swedish component of the ICLE corpus contains many more instances of modals such as *will, would, have to, should* and *might* than the LOCNESS corpus does. This is a fact about these two corpora only. Most researchers, including Aijmer, however, would wish to go further and claim that, on the whole, Swedish learners of English use more of those modals in expository writing than native English speakers do. In making such a claim, they are assuming that the Swedish component of the ICLE corpus, and the LOCNESS corpus, are both representative of the written English of native speakers of Swedish and of English respectively. It is worth considering what would make such a claim invalid. If all the essays in the Swedish corpus were written by one or two learners, for example, it would increase the chances that the evidence relates only to the idiolect of a few learners rather than to Swedish learners as a whole. Alternatively, if all the essays in one corpus (but not the other) were on the same topic and if that topic necessitated a high use of *will, would* and so on, then the judgement that the corpus accurately represented modal use by the two groups would again be thrown into doubt. If, however, each corpus can be shown to include a range of topics and a range of writers, confidence that it is representative is increased. Two further points need to be made here. Firstly, if a range of topics and writers is to be included in the corpus, it must be of a sufficient size to allow this. Thus, representativeness and size are connected. Secondly, the figures for modal use (or use of

any other feature) are averaged across the corpus/corpora. They do not take into account possible differences between learners in the same cohort. Thus, whereas we might say with confidence that Swedish learners as a group tend to use a lot of modals such as *will, would* and so on, we cannot say that every individual learner will do so. The corpus is representative of the group, not of the individual.

Finally, of course, it is not possible to extrapolate with certainty from the ICLE corpus, which contains examples of written expository English, to learners' use of a language feature in other written registers, such as narrative, or in speech, or to more advanced learners. It may be reasonable to hypothesise that Swedish learners will use the relevant modals more than native English speakers whether they are writing or speaking, and whatever kind of writing and speaking they are doing, but this remains a hypothesis until tested. On the other hand, teachers of English in Sweden may well decide that they have enough evidence to start devising ways of encouraging their learners to adopt alternative strategies to using modals, that is, that the ICLE corpus is for practical purposes representative of Swedish use of English in general.

Similar issues can be raised about any specialised corpus, but the question of representativeness really becomes controversial when applied to a general corpus, that is, one that aims to represent (a variety of) a language as a whole. There is widespread agreement that such a corpus should include texts from as many different categories of writing and speech as resources will allow. The categories are likely to include: topic areas (books and magazines on various subjects, both fiction and non-fiction, for instance); modes of publication (books, newspapers, leaflets, for example, as well as unpublished materials such as letters and diaries); social situation (casual conversation, service encounters, interviews, lessons, for example); and interactivity (monologue, dialogue and multi-party conversation). Corpora of spoken language often use standard social categories such as age, sex, socio-economic class and region to identify the different groups of people whose speech they wish to include (Leech/Rayson/Wilson 2001, 2−4). A second criterion of representativeness is that the quantity of text from a given category in the corpus would reflect its significance in the society that the corpus is to represent. For example, if twice as many books are published each year on 'social sciences' as on 'world affairs', the corpus might include twice as much text from the former as from the latter. If 15% of a population are over 60 years of age, the same proportion of the spoken component of the corpus should comprise speakers of that age.

There are, of course, considerable problems with this ideal of representativeness. One is that it is not possible to identify a complete list of 'categories' that would exhaustively account for all the texts produced in a given language. No list of domains, or genres, or social groupings can ever be complete, and indeed most general corpora explicitly exclude very specialised kinds of discourse. The ICE corpus, for example, does not include written legal discourse (Meyer 2002, 36). Those categories that are identifiable may in fact be far from homogeneous. One example is the category of 'academic discourse', which forms one of the registers used in Biber et al. (1999), but which is a composite of several different genres and many subject areas, all of which can be demonstrated to have different linguistic characteristics. Thus the 'coverage' apparently afforded by the presence of various categories may be illusory. The question of proportions is even more vexed. Gellerstam (1992, 154), for example, points out that the composition of a corpus will be very different depending on whether it is based on the amount of each kind of language that is produced or on the amount of each kind of language that most people

come into contact with. He gives the example of parliamentary proceedings, which are produced in large quantities but read by few people. In other words, there is no true measure of the 'significance' of a type of discourse to a community. Even where the ideal proportions seem to be obvious, there may be several complicating factors. Meyer (2002, 48–49), for example, reports that achieving representation of gender in a spoken corpus is a complex matter, because it is not sufficient to have equal numbers of men and women speakers, representing the broadly equal numbers in the society. As it is known that men and women tend to speak differently depending on whether they are speaking to men or women, and in what situation, a truly representative corpus needs to have equal proportions of male speakers talking to male and female addressees, in single sex and mixed groups, and the same for female speakers. What appears at first to be a simple binary distinction in fact involves at least half a dozen situational configurations, and those configurations change with each speaker change in a conversation.

There are three possible responses to the problems posed by the notion of representativeness in a general corpus. One is to avoid the notion of representation altogether, and to treat the corpus as a collection of different registers, each of which occurs frequently in the target community, but without claiming comprehensive coverage. Biber's work on register variation, for example, selects registers without claiming that together they make up a representation of English. A second is to acknowledge the problems but to do the best that is possible in the circumstances and to be transparent about how the corpus has been designed and what is in it. This allows the degree of representativeness to be assessed by the corpus user. For example, Leech/Rayson/Wilson (2001, 3) record the percentage of speech produced by speakers of different age groups in their corpus of British English conversation; the user can then decide whether this accurately reflects the distribution of ages in Britain. A third alternative is to seek to include texts from as many different sources as possible in the corpus but to treat the resulting corpus as a collection of sub-corpora rather than as a single entity. This is feasible only when each sub-corpus is of a reasonable size. The principle might be illustrated by considering the written and spoken components respectively of the British National Corpus. The spoken component comprises only 10% of the whole, which is clearly not representative either of production or of reception, but which is explained by the heavy resources required to collect spoken data in electronic form. However, 10% of 100 million words is a corpus of a respectable size that allows research to be carried out into spoken British English. The process of normalisation is used to allow valid comparisons between the written and the spoken components (Leech/Rayson/Wilson 2001). The problem of a lack of representativeness disappears. The principle of allowing size to compensate for other issues might be applied to a difficulty raised by Gellerstam (1992, 154), ironically a difficulty caused by representativeness. Gellerstam notes that 75% of all written output in Swedish in any given year comprises newspaper texts. A corpus of written Swedish that consisted of this proportion of newspaper text, and so was representative in that sense, would include only small amounts of other kinds of Swedish. A user of the corpus would be in danger of seeing nothing but the newspaper texts. This would be true, unless the corpus as a whole was large enough so that even 25% of the total could comprise hundreds of millions of words, and unless it was possible to access the various components of the corpus independently. In that case, non-journalistic Swedish could be investigated and compared with the newspaper texts when required.

7. The corpus as artefact

The second issue often discussed in terms of corpus design is balance. Balance refers to the internal composition of the corpus, that is to the proportions of the various sub-corpora that make it up. A corpus that consists of much more of one kind of text than another may be said to be unbalanced. It is immediately obvious, as illustrated by Gellerstam's discussion of Swedish corpora, that balance (equality between sub-corpora) may be at odds with representativeness (each sub-corpus in proportion to its significance). An issue that illustrates this is the decision that the corpus-builder has to take between 'number of texts' and 'number of tokens'. A hypothetical researcher may wish to study newspaper editorials. Obviously, a balanced corpus would consist of the same amount of text from each of the newspapers concerned. However, if some newspapers typically print more, or longer, editorials than others, the problem of 'sameness' arises. The corpus builder could select the same number of editorials from each newspaper, in which case the sub-corpora would contain unequal numbers of words because some texts are longer than others. On the other hand, balancing the sub-corpora in terms of tokens would lead to inequality in terms of number of texts, and would in addition be unrepresentative of the balance of text actually produced by the newspapers concerned.

An example of a corpus designed to be balanced is the Michigan corpus of Spoken Academic English, or MICASE. Its contents are subdivided by speech event type (lecture, discussion group, seminar, meetings, office hours, service encounters etc.), by academic subject (physical sciences, social sciences, humanities etc.), by 'participant level' (undergraduate student, postgraduate student, junior or senior faculty etc.), and by primary discourse mode (monologue, discussion etc.). The corpus is also divisible by the attributes of the speakers: their age, sex, first language, and so on. An argument for the balance of the corpus is that each set of subdivisions is roughly equal. For example: 46% of the speech is produced by men, 54% by women; 49% is produced by faculty, 44% by students; and approximately a quarter of the content of the corpus comes from each of the four main academic subject areas (the actual figures are between 19% and 26%). However, only 12% of the speech is from non-native speakers of English, and only 8% belongs to the discourse mode classified as 'panel' (where a group of speakers each produces a short monologue in turn). These figures presumably reflect the relatively low proportion of non-native speakers of English on the Michigan campus and the relative infrequency of panel-type discourse. Furthermore, only 14% of the speech in the 'monologic' mode is produced by students, with the rest being produced by faculty. Although overall male/female proportions are roughly equal, the numbers are not equal in all the subject areas. The biggest discrepancy is in Social Sciences and Education, where 63% of the speech is produced by female speakers. Again, this is no doubt an inevitable consequence of the contexts of recording: few lectures are given by students and there are sex imbalances in some academic disciplines. Although the corpus designers planned MICASE as a balanced, rather than a representative, corpus, lack of balance in the context has affected the corpus to some extent.

Although the need for balance in a corpus may appear obvious, it is worth considering precisely what benefits a balanced corpus offers. To take an example: a researcher looking at a particular language feature in monologic discourse in MICASE does not need to worry that the prevalence or absence of the feature is due to the peculiarities of male or female speech rather than to the nature of monologue, because the

monologic discourse component of the corpus is split exactly 50:50 between men and women. If it is found that the feature occurs more frequently in monologues by women than in monologues by men, on the other hand, this can confidently be ascribed to a difference in gendered speech rather than to a difference in the proportions of men and women in that component of the corpus. It should be noted, however, that a similar effect can be obtained by normalising the frequency of an item, so that differences between the sizes of sub-corpora are overcome. When Poos/Simpson (2002) normalise the frequency of hedges such as *kind of* and *sort of* in MICASE monologues, they find that the frequency per thousand words is practically the same for men and women. They express greater confidence in this finding for those disciplines where the quantity of male and female speech in monologues is approximately the same than in those where it is markedly different. In the Physical Sciences, for example, although the per thousand word frequency is almost the same for men and women, there are only two female speakers, and the number of words produced by them is less than half that produced by the male speakers (Poos/Simpson 2002, 8). It is possible, then, that one or two of the female speakers have used an abnormal number of hedges, so skewing the results. In other words, the authors argue that a lack of balance may lead to a lack of representativeness, in that women producing monologues in the Physical Sciences are under-represented in this corpus.

This suggests that the real benefit of a balanced corpus is that each of its various components is large enough to make comparisons feasible. In MICASE there are sufficient quantities of male and female speech, of speech by students and faculty, of monologic and interactive speech, and of speech in the various subject areas, to warrant comparisons between these categories, even though the total word count of the corpus (about 1.5 million words) is not huge. Another point to be made is that balance, like representativeness, implies explicitness in corpus description. Poos/Simpson (2002, 7) point out that an imbalance in a corpus does not matter so long as it is known and hypotheses can be adjusted accordingly. They note, for instance, that in MICASE there is more interactive speech, proportional to monologue, in Biology than in Social Sciences. If a language feature were found to be more prevalent in Biology than in Social Sciences, then, a possible explanation would be that this feature was frequent in interactive speech generally. This would have to be explored before an association between the feature and discipline was proposed.

The explicitness of description, however, can only be partial. In fact, arguments that a particular corpus is representative, or balanced, are inevitably circular, in that the categories we are invited to observe are artefacts of the design procedure. The categories that have formed the basis of the corpus design are, indeed, representative or balanced, but other categories may be less representative or balanced and less observable. Meyer's discussion of gender balance is an example of this. If the only categories considered are 'male' versus 'female', then a corpus may be designed to capture an equal amount of speech from men and women, and the results may show that this has been achieved. Other categories, such as 'addressee', that have not been built into the corpus design, may be in the end very unbalanced. The women may have chosen to make recordings only when speaking to other women, for example, or the men may have avoided recording all-male chat.

8. Large and small corpora

In looking at the issue of corpus size, we are once again faced with a paradox. On the one hand, it might said of any corpus that the larger it is, the better, the only upper constraint being computational capacity and speed of software (Sinclair 1991, 18; Meyer 2002, 33). As has been indicated above, some of the difficulties posed by seeking to make a corpus balanced and representative can be lessened by having a corpus large enough for each of its constituent components to be of a substantial size (cf. Aston/Burnard 1998, 21). The only advantage of a small corpus is that the occurrence of very frequent words is low enough to make observation of all instances feasible, whereas in a large corpus some kind of sampling has to take place (Carter/McCarthy 1995, 143). The counter-argument is that such sampling can incorporate the observation of large-scale patterning rather than simply taking a small sub-set of the whole. For example, collocation lists can be used to summarise the information from a large number of concordance lines so that a smaller number of lines incorporating more specific phraseologies can then be examined in detail. This is true even of very frequent grammatical items, if these are considered to be the locus of phraseology rather than simply a grammatical category. For example, Groom (2007) investigates the behaviour of very frequent items, such as prepositions, in corpora of academic writing distinguished by discipline. He notes certain phraseologies that are typical of one discipline rather than another, such as the sequence 'It is in … that' which is an identifying phraseology of Literary Criticism and which functions as an introduction to an interpretative observation (e. g. *it is in the exchanges between these characters that Shakespeare can again emphasise the political ambiguity of language*). This sequence was identified in the course of an investigation of the very frequent word *in*. Sequences such as this one are potentially so long, however, that even Groom's Literary Criticism corpus of 4 million words, with its many thousands of instances of *in*, cannot always show more than a handful of each.

It would appear, then, that any corpus should simply be as large as possible, and that to achieve this the corpus should continue being added to over the lifetime of its use. In most situations, however, this is impractical. The need to plan the resources involved in a research project, including the time and money involved, make it necessary to specify in advance the size of the corpus to be compiled. If a corpus is extensively annotated, and especially if some of this annotation has to be done or edited manually, increasing the size of the corpus greatly increases the amount of effort involved. Adding to the corpus once the resources allocated to its compilation have been used up is impossible. Even if such problems do not exist, if a corpus is converted to digits for storage, enlarging the corpus means re-converting the whole entity. As a result, additions cannot be made with great frequency.

Connected to the issue of corpus size is that of sample size. A corpus of a million words or so cannot afford to include whole books which might be up to 100,000 words in length, and as a result text sampling is often used. The British component of the International Corpus of English, for example, consists of 'texts' of 2,000 words each (Nelson/Wallis/Aarts 2002, 4). Each 'text' consists either of part of a longer entity, such as a novel, or of a collection of smaller entities, such as business letters. Such uniformity ensures maximum control over the content of the corpus, which is advantageous in a situation where corpora (in English, from around the world) are to be compared. Sinclair (1991, 19), however, argues that sampling can lead to differences between parts of a text

being overlooked. The 'whole text' policy that he advocates, however, does necessitate the collection of a much larger corpus if one or two large publications are not to affect the output disproportionately.

In the long run, then, the issue of corpus size becomes a set of interconnecting issues that concern the aims and methods of investigation as well as the question of size. Where resources are limited, or where close control is needed to ensure comparability between corpora, or where a very accurate transcription or extensive annotation is required, the corpus will tend to be relatively small, and, if it is a general corpus, will probably consist of samples of texts rather than whole texts. If size and whole texts are seen as priorities, annotation is likely to be minimal and comparability, even between sections of the corpus, is unlikely to be exact. Conversely, a small corpus is most useful if it is annotated, and in turn an annotated corpus is most useful for investigating the relative frequency and other aspects of instances of the categories for which it has been annotated. For example, Semino and Short's study of speech, thought and writing representation in newspapers, novels and biographies is based on a corpus of just under 260,000 words that is, nonetheless, minutely annotated (Semino/Short 2004, 19). A large corpus is most useful for studying less frequent items or, crucially, the macro-patterning of language that is not amenable to intuition and is ignored by grammatical tradition, and that can only be seen when many instances of relatively long sequences of items are brought together.

9. Conclusion

Corpus design and compilation seems like a simple matter. The researcher decides what the various components of the corpus are to consist of and what the size relationship between the components will be. He or she then identifies places where the desired texts can be found, and so builds the corpus. In practice, the situation is likely to be much more complex. Practical issues such as copyright restrictions, or availability in electronic form, may determine which texts are used and as a consequence which variables can be taken into account. In considering the relative size of components, the researcher may need to choose between balance and representativeness. What would constitute 'representation' may not be identifiable anyway.

In addition, although in theory a corpus is a neutral resource that can be used in research from any number of standpoints (Leech 1997, 7), in practice the design of the corpus may strongly constrain the kind of research that is carried out (Sinclair 1992). Most obviously, a corpus that is limited in time-period precludes discourse studies that depend on a diachronic dimension to intertextuality (Teubert 2003, 12). A corpus that is small and balanced prioritises the investigation of variation using grammatical categories (Biber et al. 1999, 15–24). A very large corpus facilitates the study of phraseology and macro-patterning (Sinclair 2004, 24–48).

Far from being neutral, then, issues of corpus design and building take us to the heart of theories of corpus linguistics. Questions of what goes into a corpus are largely answered by the specific research project the corpus is designed for, but are also connected to more philosophical issues around what, potentially, corpora can show us about language.

10. Literature

Aijmer, K. (2002), Modality in Advanced Swedish Learners' Written Interlanguage. In: Granger, S./Hung, J./Petch-Tyson, S. (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: Benjamins, 55–76.

Aston, G./Burnard, L. (1998), *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baldry, A./Thibault, P. (2005), Multimodal Corpus Linguistics. In: Thompson, G./Hunston, S. (eds.), *System and Corpus: Exploring Connections*. London: Equinox, 164–183.

Barr, G. K. (2003), Two Styles in the New Testament Epistles. In: *Literary and Linguistic Computing* 18, 235–248.

Bernardini, S./Baroni, M. (2004), Web-mining Disposable Corpora in the Translation Classroom. Paper read at the 6th TALC Conference, Granada, 2004.

Biber, D. (1988), *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D./Finegan, E. (2001), Diachronic Relations among Speech-based and written registers in English. In: Conrad, S./Biber, D. (eds.), Variation in English: Multidimensional Studies. Harlow etc.: Longman, 66–83.

Biber, D./Finegan, E./Atkinson, D. (1994), ARCHER and its Challenges: Compiling and Exploring a Representative Corpus of Historical English Registers. In: Fries, U./Tottie, G./Schneider, P. (eds.), *Creating and Using English Language Corpora: Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora, Zürich 1993*. Amsterdam: Rodopi, 1–14.

Biber, D./Johansson, S./Leech, G./Conrad, S./Finegan, E. (1999), *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Carter, R./McCarthy, M. (1995), Grammar and the Spoken Language. In: *Applied Linguistics* 16(2), 141–158.

Du Bois, J./Chafe, W./Meyer, C./Thompson, S. (2000), *Santa Barbara Corpus of Spoken American English, Part 1*. Philadelphia: Linguistic Data Consortium.

Du Bois, J./Chafe, W./Meyer, C./Thompson, S. (2003), *Santa Barbara Corpus of Spoken American English, Part 2*. Philadelphia: Linguistic Data Consortium.

Gellerstam, M. (1992), Modern Swedish Text Corpora. In: Svartvik, J. (ed.), *Directions in Corpus Linguistics*. Berlin/New York: Mouton de Gruyter, 149–163.

Granger, S. (1998), The Computer Learner Corpus: A Versatile New Source of Data for SLA Research. In: Granger, S. (ed.), *Learner English on Computer*. London: Longman, 3–18.

Groom, N. (2007), Phraseology and Epistemology in Humanities Writing: A Corpus-driven Study. PhD thesis, University of Birmingham.

Francis, N./Kučera, H. (1982), *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Leech, G. (1997), Introducing Corpus Annotation. In: Garside, R./Leech, G./McEnery, A. (eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman, 1–18.

Leech, G./Rayson, P./Wilson, A. (2001), *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Longman.

Meyer, C. (2002), *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Nelson, G./Wallis, S./Aarts, B. (2002), *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.

Poos, D./Simpson, R. (2002), Cross-disciplinary Comparisons of Hedging: Some Findings from the Michigan Corpus of Academic Spoken English. In: Reppen, R./Fitzmaurice, S./Biber, D. (eds.), *Using Corpora to Explore Linguistic Variation*. Amsterdam: Benjamins, 3–23.

Renouf, A. (1987), Corpus Development. In: Sinclair, J. (ed.), *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: HarperCollins, 1–40.

Rottweiler, G. (2006), Evaluative Meanings, Social Values, and One Lexical Set: A Corpus Analysis of *Unilateral, Unilaterally, Unilateralism*, and *Unilateralist/s*. MPhil thesis, University of Birmingham.
Scott, M. (2004), *WordSmith Tools*, version 4.0. Oxford: Oxford University Press.
Semino, E./Short, M. (2004), *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. London: Routledge.
Sinclair, J. (1991), *Corpus, Concordance Collocation*. Oxford: Oxford University Press.
Sinclair, J. (1992), The Automatic Analysis of Corpora. In: Svartvik, J. (ed.), *Directions in Corpus Linguistics*. Berlin/New York: Mouton de Gruyter, 379–397.
Sinclair, J. (2004), *Trust the Text: Language, Corpus and Discourse*. London: Routledge.
Stubbs, M. (1996), *Text and Corpus Analysis:* Computer-Assisted Studies of Language and Culture. Oxford: Blackwell.
Teubert W. (2001), A Province of a Federal Superstate, Ruled by an Unelected Bureaucracy: Keywords of the Euro-sceptic Discourse in Britain. In: Musolff, A./Good, C./Points, P./Wittlinger, R. (eds.), *Attitudes Towards Europe*. Aldershot: Ashgate, 45–86.
Teubert W. (2003), Writing, hermeneutics, and corpus linguistics. In: *Logos and Language* 4, 1–17.
Teubert, W. (2004), When did we Start Feeling Guilty? In: Weigand, E. (ed.), *Emotion in Dialogic Interaction*. Amsterdam: Benjamins, 121–162.

Susan Hunston, Birmingham (UK)

10. Text corpora

1. Introduction

Among corpora, one often distinguishes between text corpora (consisting of written material – both published and unpublished), spoken corpora (cf. article 11) and multimodal corpora (cf. article 12). More specialized types of corpora are treebanks (cf. article 13), historical, learner, or parallel corpora (cf. articles 14, 15 and 16, respectively). On closer inspection, the distinction between text and speech corpora is not as straightforward as it appears to be. The article will therefore begin with a discussion of the notions 'text' and 'speech'. Section 1.1.1. will consider how far text corpora can be defined as collections of written texts and what it means for a text to be 'written'. Corpora further have to be distinguished from mere text databases or text archives (section 1.1.2.). An outline of topics treated in the remaining sections of this article will be given in section 1.2.