ahds literature, languages and linguistics

Developing Linguistic Corpora: a Guide to Good Practice

Corpus and Text — Basic Principles

John Sinclair, Tuscan Word Centre

A corpus is a remarkable thing, not so much because it is a collection of language text, but because of the properties that it acquires if it is well-designed and carefully-constructed.

The guiding principles that relate corpus and text are concepts that are not strictly definable, but rely heavily on the good sense and clear thinking of the people involved, and feedback from a consensus of users. However unsteady is the notion of *representativeness*, it is an unavoidable one in corpus design, and others such as *sample* and *balance* need to be faced as well. It is probably time for linguists to be less squeamish about matters which most scientists take completely for granted.

I propose to defer offering a definition of a corpus until after these issues have been aired, so that the definition, when it comes, rests on as stable foundations as possible. For this reason, the definition of a corpus will come at the end of this paper, rather than at the beginning.

1. Who builds a corpus?

Experts in corpus analysis are not necessarily good at building the corpora they analyse — in fact there is a danger of a vicious circle arising if they construct a corpus to reflect what they already know or can guess about its linguistic detail. Ideally a corpus should be designed and built by an expert in the communicative patterns of the communities who use the language that the corpus will mirror. Quite regardless of what is inside the documents and speech events, they should be selected as the sorts of documents that people are writing and reading, and the sorts of conversations they are having. Factual

Notes

evidence such as audience size or circulation size can refine such sampling. The corpus analyst then accepts whatever is selected.

This could be stated as a principle:

1. The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise.

Obviously if it is already known that certain text types contain large numbers of a microlinguistic feature such as proper nouns or passive verb phrases, it becomes a futile activity to "discover" this by assembling a corpus of such texts.

Selection criteria that are derived from an examination of the communicative function of a text are called *external criteria*, and those that reflect details of the language of the text are called *internal criteria*. Corpora should be designed and constructed exclusively on external criteria (Clear 1992)[1].

2. What is a corpus for?

A corpus is made for the study of language; other collections of language are made for other purposes. So a well-designed corpus will reflect this purpose. The contents of the corpus should be chosen to support the purpose, and therefore in some sense represent the language from which they are chosen.

Since electronic corpora became possible, linguists have been overburdened by truisms about the relation between a corpus and a language, arguments which are as irrelevant as they are undeniably correct. Everyone seems to accept that no limits can be placed on a natural language, as to the size of its vocabulary, the range of its meaningful structures, the variety of its realisations and the evolutionary processes within it and outside it that cause it to develop continuously. Therefore no corpus, no matter how large, how carefully designed, can have exactly the same characteristics as the language itself.

Fine. So we sample, like all the other scholars who study unlimitable phenomena. We

remain, as they do, aware that the corpus may not capture all the patterns of the language, nor represent them in precisely the correct proportions. In fact there are no such things as "correct proportions" of components of an unlimited population.

2. Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen.

However hard we strive, a corpus will occasionally show features which we suspect not to be characteristic of the language under study, or fail to show features which are expected. Following our first principle above, we should not feel under pressure to use the patterns of the language to influence the design of the corpus, but we should review the design criteria to check that they are adequate.

To optimise the application of this principle we can make use of an important resource within ourselves, which is not available to most scientific researchers in other disciplines. As sophisticated users of at least one language, we have an inbuilt awareness of language structure, often called intuition, that gives a personal, independent and non-negotiable assessment of language pattern. Intuition can help in many ways in language research, in conjunction with other criteria of a more examinable nature. The drawbacks to intuition are (a) that we cannot justify its use beyond personal testimony, and (b) that people differ notoriously in their intuitive judgements. In this context we should also be aware that an incautious use of intuition in the selection of texts for a corpus would undermine the first principle[2].

3. How do we sample a language for a corpus?

There are three considerations that we must attend to in deciding a sampling policy:

1. The orientation to the language or variety to be sampled.
2. The criteria on which we will choose samples.
3. The nature and dimensions of the samples.

1. Orientation

This is not a crisply delineated topic, and has largely been taken for granted so far in corpus building. The early corpora, for example the Brown corpus and those made on its model (Hofland and Johansson 1982), were *normative* in their aims, in that their designers wanted to find out about something close to a standard language. The word "standard" appears in the original Brown title; by choosing published work only, they automatically deselected most marked varieties. Most of the large reference corpora of more recent times adopt a similar policy; they are all constructed so that the different components are like facets of a central, unified whole. Such corpora avoid extremes of variation as far as possible, so that most of the examples of usage that can be taken from them can be used as models for other users.

Some corpora have a major variable already as part of the design — a historical corpus, for example, is deliberately constructed to be internally contrastive, not to present a unified picture of the language over time (though that could be an interesting project). Another kind of corpus that incorporates a time dimension is the *monitor* corpus (Sinclair 1982); a monitor corpus gathers the same kind of language at regular intervals and its software records changes of vocabulary and phraseology. *Parallel* corpora, or any involving more than one language, are of the same kind — with inbuilt contrasting components; so also is the small corpus used in Biber et. al. (1999) to demonstrate varietal differences among four externally-identified varieties of contemporary English. These corpora could be called *contrastive* corpora because the essential motivation for building them is to contrast the principal components.

There is a guiding principle here of great importance, and one which is commonly ignored.

3. Only those components of corpora which have been designed to be independently contrastive should be contrasted.

That is to say, the existence of components differentiated according to the criteria discussed below, or identified by archival

information, does not confer representative status on them, and so it is unsafe to use them in contrast with other components. Now that with many corpus management systems it is possible to "dial-a-corpus" to your own requirements, it is important to note that the burden of demonstrating representativeness lies with the user of such selections and not with the original corpus builder. It is perfectly possible, and indeed very likely, that a corpus component can be adequate for representing its variety within a large normative corpus, but inadequate to represent its variety when freestanding.

This point cannot be overstated; a lot of research claims authenticity by using selections from corpora of recognised standing, such as the Helsinki Corpus, which is a notable reference corpus covering the language of almost a millennium in a mere 1,572,820 words. Each small individual component of such a corpus makes its contribution to the whole and its contrasts with other segments, but was never intended to be a freestanding representative of a particular state of the language. See the detailed description at http://icame.uib.no/hc/. Normative, historical, monitor and varietal corpora are not the only kinds; demographic sampling has been used a little, and there are all sorts of specialised corpora. For an outline typology of corpus and text see Sinclair (2003), which is a summary and an update of a report made for the European Commission (for that report see the EAGLES server at http://www.ilc.pi.cnr.it).

2. Criteria

Any selection must be made on some criteria and the first major step in corpus building is the determination of the criteria on which the texts that form the corpus will be selected.

Common criteria include:

a. the mode of the text; whether the language originates in speech or writing, or perhaps nowadays in electronic mode;
b. the type of text; for example if written, whether a book, a journal, a notice or a letter;
c. the domain of the text; for example whether academic or popular;

d. the language or languages or language varieties of the corpus;
e. the location of the texts; for example (the English of) UK or Australia;
f. the date of the texts.

Often some of these large-scale criteria are pre-determined by constraints on the corpus design — for example a corpus called MICASE stands for the Michigan Corpus of Academic Spoken English, and the corpus consists of speech events recorded on the Ann Arbor campus of the University of Michigan on either side of the millennium; it follows that the language in the corpus will mainly be of the large variety called American English. All the above criteria are pre-determined, and all but the date are built into the name of this corpus, so its own structural criteria will be set at a more detailed level[3].

All but the most comprehensive corpora are likely to use one or more criteria which are specific to the kind of language that is being gathered, and it is not possible to anticipate what these are going to be. The corpus designer should choose criteria that are easy to establish, to avoid a lot of labour at the selection stage, and they should be of a fairly simple kind, so that the margin of error is likely to be small. If they are difficult to establish, complex or overlapping they should be rejected, because errors in classification can invalidate even large research projects and important findings.

Now that there are a number of corpora of all kinds available, it is helpful to look at the criteria that have been used, and to evaluate them in three ways — as themselves, how useful and valuable a variety of the language they depict; as a set of criteria, how they interact with each other and avoid ambiguity and overlap; and the results that they give when applied to the corpus.

4. Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination.

Beyond these criteria it is possible to envisage an unlimited categorisation of people, places and events, any of which are potentially

valuable for one study or another (see the typology mentioned above). The gender of the originator of a text has been a popular criterion in recent years, though few texts have a single originator whose gender is known, and hoaxes are not unknown (for example it was recently alleged that the works of a famous crime writer noted for rough-and-tough stories were in fact composed by his wife). It is essential in practice to distinguish structural criteria from useful information about a text.

For a corpus to be trusted, the structural criteria must be chosen with care, because the concerns of balance and representativeness depend on these choices. Other information about a text can, of course, be stored for future reference, and scholars can make up their own collections of texts to suit the objectives of their study. The question arises as to how and where this information should be stored, and how it should be made available. Because it is quite commonly added to the texts themselves, it is an issue of good practice, especially since in some cases the additions can be much larger than the original texts.

In the early days of archiving text material, the limitations of the computers and their software required a structurally simple model; also before there was an abundance of language in electronic form, and before the internet made it possible for corpora to be accessed remotely, it was necessary to agree protocols and practices so that data could be made available to the research community. The model that gained widest acceptance was one where additional material was interspersed in the running text, but enclosed in diamond brackets so that it could — at least in theory — be found quickly, and ignored if the text was required without the additions.

Nowadays there is no need to maintain a single data stream; modern computers have no difficulty storing the plain text without any additions, and relating it token by token to any other information set that is available, whether "mark-up", which is information about the provenance, typography and layout of a printed document, or "annotation", which is analytic information usually about the language[4]. It is also possible nowadays to

store facsimiles of documents and digitised recordings of speech, and have the computer link these, item by item, to plain text, thus removing even the need to have mark-up at all.

5. Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications.

3. Sampling

Looking down from the totality of the corpus, the major criteria will define several *components*, while at the other end are the individual *texts*, which will be such things as written or printed documents, and transcripts of spoken events. *Cells* are the groupings formed from the intersection of criteria.

The first-level components will be small in number, for practical reasons, because if there are too many then either each component will be very small or the corpus will be very large. The simplest classification is binary, so that if a corpus of spoken language is first divided into "private" and "public", then each of these types will have to be represented by a sufficiently large amount of text for its characteristics to become evident. If the next criterion is "three or fewer active participants", as against "more than three active participants", then each of the original categories is divided into two, and the theoretical size of the corpus doubles.

Each criterion divides the corpus into smaller cells; if we assume that the criteria are binary and cross-cutting then (as we have just seen) two criteria divide the corpus into four cells, three into eight, four into sixteen etc. You then have to decide what is the acceptable minimum number of words in a cell; this depends quite a lot on the type of study you are setting out to do, but if it is not substantial then it will not supply enough reliable evidence as part of the overall picture that the corpus gives of the language. This is known as the "scarce data problem". The matter of size is discussed later, and the example in the following paragraph is only illustrative.

If you decide on, say, a million words as the minimum for a cell, then with four criteria you

need a corpus with a minimum size of sixteen million words. Each additional binary criterion doubles the minimum size of the corpus, and in addition we find that real life is rarely as tidy as this model suggests; a corpus where the smallest cell contains a million words is likely in practice to have several cells which contain much more. This involves the question of balance, to which we will return. There are also questions of criteria that have more than two options, and of what to do with empty or underfilled cells, all of which complicate the picture.

The matter of balance returns as we approach the smallest item in a corpus, the text. Here arises another issue in sampling that affects, and is affected by, the overall size of the corpus. Language artefacts differ enormously in size, from a few words to millions, and ideally, documents and transcripts of verbal encounters should be included in their entirety. The problem is that long texts in a small corpus could exert an undue influence on the results of queries, and yet it is not good practice to select only part of a complete artefact. However it is an unsafe assumption that any part of a document or conversation is representative of the whole — the result of research for decades of discourse and text analysis make it plain that position in a communicative event affects the local choices.

The best answer to this dilemma is to build a large enough corpus to dilute even the longest texts in it. If this is not practical, and there is a risk that a single long text would have too great an influence on the whole, so recourse has to be made to selecting only a part of it, and this has to be done on "best guess" grounds. But even a very large corpus may find it almost impossible to get round copyright problems if the builders insist on only complete texts. The rights holders of a valuable document may not agree to donate the full text to a corpus, but if it is agreed that occasional passages are omitted, so that the value of the document is seriously diminished, then the rights holders might be persuaded to relent.

These are the issues in text selection, and one point in particular should be made clearly. There is no virtue from a linguistic point of view in selecting samples all of the same size. True, this was the convention in some of the

early corpora, and it has been perpetuated in later corpora with a view to simplifying aspects of contrastive research. Apart from this very specialised consideration, it is difficult to justify the continuation of the practice. The integrity and representativeness of complete artefacts is far more important than the difficulty of reconciling texts of different dimensions.

6. Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size.

4. Representativeness

It is now possible to approach the notion of representativeness, and to discuss this concept we return to the first principle, and consider the users of the language we wish to represent. What sort of documents do they write and read, and what sort of spoken encounters do they have? How can we allow for the relative popularity of some publications over others, and the difference in attention given to different publications? How do we allow for the unavoidable influence of practicalities such as the relative ease of acquiring public printed language, e-mails and web pages as compared with the labour and expense of recording and transcribing private conversations or acquiring and keying personal handwritten correspondence? How do we identify the instances of language that are influential as models for the population, and therefore might be weighted more heavily than the rest?

The previous paragraph is a set of questions to which there are no definite answers, and yet on which the whole character of the corpus will rest. According to claims, the most likely document that an ordinary English citizen will cast his or her eye over is *The Sun* newspaper; in a corpus of British English should we then include more texts from that paper than from any other source? If this argument is rejected on stylistic grounds — perhaps that the language of *The Sun* is particularly suited to the dramatic presentation of popular news and views and

would not be recommended as a general model for written work — then the corpus builder is adopting a prescriptive stance and is risking the vicious circle that could so easily arise, of a corpus constructed in the image of the builder.

The important steps towards achieving as representative a corpus as possible are:

a. decide on the structural criteria that you will use to build the corpus, and apply then to create a framework for the principal corpus components;
b. for each component draw up a comprehensive inventory of text types that are found there, using external criteria only;
c. put the text types in a priority order, taking into account all the factors that you think might increase or decrease the importance of a text type — the kind of factors discussed above;
d. estimate a target size for each text type, relating together (i) the overall target size for the component (ii) the number of text types (iii) the importance of each (iv) the practicality of gathering quantities of it;
e. as the corpus takes shape, maintain comparison between the actual dimensions of the material and the original plan;
f. (most important of all) document these steps so that users can have a reference point if they get unexpected results, and that improvements can be made on the basis of experience.

Let me give one simple example of these precepts in operation. The precursor of The Bank of English contained a substantial proportion of the quality fiction of the day. This came from a request from one of the sponsors, who felt that a corpus was such an odd thing (in 1980 it *was* an odd thing) that users of the end products would be reassured if there was quite a lot of "good writing" in it. That is to say, under (a) above it was decided that there should be emphasis on this kind of writing; this decision affected the choice of texts under (b) also. However, one of the main aims of creating the corpus was to retrieve evidence in support of the learning of the English language, and the requirements of this

mundane purpose clashed with some of the prominent features of modern fiction. For example, the broad range of verbs used to introduce speech in novels came out rather too strongly — *wail*, *bark* and *grin* are all attested in this grammatical function, and while their occurrence is of interest to students of literary style, they are of limited utility to learners seeking basic fluency in English (Sinclair et. al. 1990 p. 318).

This clash between the design of the corpus and its function became clear as soon as work started on the first Cobuild grammar (1990). Because the corpus had been carefully designed and fully documented, it was possible to determine — and therefore roughly counterbalance — the bias that had been introduced. In fairness to the original designers, it should be emphasised that there were no previous models to turn to at that time, and no way of assessing the effects of different varieties of a language.

A corpus that sets out to represent a language or a variety of a language cannot predict what queries will be made of it, so users must be able to refer to its make-up in order to interpret results accurately. In everything to do with criteria, this point about documentation is crucial. So many of our decisions are subjective that it is essential that a user can inspect not only the contents of a corpus but the reasons that the contents are as they are. Sociolinguistics is extremely fashion-conscious, and arguments that are acceptable criteria during one decade may look very old-fashioned in the next.

Also at any time a researcher may get strange results, counter-intuitive and conflicting with established descriptions. Neither of these factors proves that there is something wrong with the corpus, because corpora are full of surprises, but they do cast doubt on the interpretation of the findings, and one of the researcher's first moves on encountering unexpected results will be to check that there is not something in the corpus architecture or the selection of texts that might account for it.

7. The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.

5. Balance

The notion of balance is even more vague than representativeness, but the word is frequently used, and clearly for many people it is meaningful and useful. Roughly, for a corpus to be pronounced balanced, the proportions of different kinds of text it contains should correspond with informed and intuitive judgements.

Most general corpora of today are badly balanced because they do not have nearly enough spoken language in them; estimates of the optimal proportion of spoken language range from 50% — the neutral option — to 90%, following a guess that most people experience many times as much speech as writing. Another factor that affects balance is the degree of specialisation of the text, because a specialised text in a general corpus can give the impression of imbalance.

This is a problem in the area of popular magazines in English, because there are a large number of them and most use a highly specialised language that non-devotees just do not understand. So as a text type it is a very important one, but it is almost impossible to select a few texts which can claim to be representative. How are magazines for fly fishermen, personal computers and popular music going to represent the whole variety of popular magazines (as is the case in The Bank of English)?

It was mentioned above that not all cells need to be filled; for example the written component of a corpus may subdivide into newspapers, magazines, books etc., for which there are no exact equivalents in the spoken language, which might divide into broadcasts, speaker-led events, organised meetings and conversations. The idea of maintaining a balance prompts the corpus builder to try to align these categories, however roughly, so that there is not too much very formal or very informal language in the corpus as a whole. If — as is frequently reported — many users value informal and impromptu language as revealing most clearly how meaning is made, a deliberate imbalance may be created by selection in favour of this variety, and this should be documented so that users are aware of the bias that has been knowingly introduced into the corpus.

Specialised corpora are constructed after some initial selectional criteria have been applied, for example the MICASE corpus cited above. More delicate criteria are used to partition them, but the issues of balance and representativeness remain cogent and central in the design.

8. The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.

6. Topic

The point above concerning a text type where most of the exemplars are highly specialised, raises the matter of topic, which most corpus builders have a strong urge to control. Many corpus projects are so determined about this that they conduct a semantic analysis of the language on abstract principles like those of Dewey or Roget, and then search for texts that match their framework. Three problems doom this kind of enterprise to failure. One is that the corpus will not conform to the classification, the second (also faced by library cataloguers) is that no two people agree on any such analysis, and the third is that topic classification turns out to be much more sociolinguistic than semantic, and therefore dependent on the culture and not on the meanings of the words. This last point emerges strongly when we try to make corpora in more than one language but sharing the same topic classification.

As well as these practical points, our first principle rules out topic as a source of corpus criteria. The most obvious manifestation of topic is certainly found in the vocabulary, and the notion of vocabulary as a defining characteristic of a corpus is strong; hence it seems strange to many people that it is essential that the vocabulary should not be directly controlled. But vocabulary choice is clearly an internal criterion.

There are external correlates, and these will indirectly control the vocabulary of the selected texts. For example many social institutions, like professional bodies and educational establishments, do show the kind

of vocabulary consistency at times that we associate with topic, and they can be used as external criteria, but topic is most definitely a matter of language patterns, mainly of vocabulary selection and discourse focusing.

9. Any control of subject matter in a corpus should be imposed by the use of external, and not internal, criteria.

7. Size

The minimum size of a corpus depends on two main factors:

1. the kind of query that is anticipated from users,
2. the methodology they use to study the data.

There is no maximum size. We will begin with the kind of figures found in general reference corpora, but the principles are the same, no matter how large or small the corpus happens to be. To relate the kind of query to the size of the corpus, it is best to start with a list of the "objects" that you intend to study; the usual objects are the physical word forms or objects created by tags, such as lemmas. Then try them out on one of the corpora that is easy to interrogate, such as the million-word corpora on the ICAME CD-ROM (Hofland 1999). The Brown-group of corpora are helpful here, because they have been proof-read and tagged and edited over many years, and with a million words the sums are easy.

To illustrate how this can be done, let us take the simple case of a researcher wishing to investigate the vocabulary of a corpus. For any corpus one of the first and simplest queries is a list of word forms, which can be organised in frequency order. (NB word forms are not lemmas, where the various inflections of a "word" in the everyday sense are gathered together, but the message would not be much different with lemmas[5].

The frequencies follow Zipf's Law (1935), which basically means that about half of them occur once only, a quarter twice only, and so on. So for the first million-word corpus of general written American English (the Brown corpus), there was a vocabulary of different word forms of 69002, of which 35065

occurred once only. At the other end of the frequency scale, the commonest word, *the* has a frequency of 69970, which is almost twice as common as the next one, *of*, at 36410.

There is very little point in studying words with one occurrence, except in specialised research, for example authorship studies (Morton 1986). Recurrence — a frequency of two or more — is the minimum to establish a case for being an independent unit of the language; but only two occurrences will tell us very little indeed about the word. At this point the researcher must fix a minimum frequency below which the word form will not be the object of study. Let us suggest some outline figures that may guide practice. A word which is not specially ambiguous will require at least twenty instances for even an outline description of its behaviour to be compiled by trained lexicographers. But there are other factors to consider, the consequences of what seems to be a general point that alternatives — members of a set or system — are often not equally likely. The same tendency that we see in Zipf's Law is found in many other places in the numerical analysis of a corpus. Very often the main meaning or use or grammatical choice of a word is many times as frequent as the next one, and so on, so that twenty occurrences may be sufficient for the principal meaning of a word, while some quite familiar senses may occur only seldom. This applies also to frequent words which can have some important meanings or uses which are much less common than the principal ones. Word classes occur in very different proportions, so if the word can be both noun and verb, the verb uses are likely to be swamped by the noun ones, and for the verb uses researchers often have recourse to a tagged corpus. In many grammatical systems one choice is nine times as common as the other (Halliday 1993), so that for every negative there are nine positives.

So some additional leeway will have to be built in to cope with such contingencies. If the objects of study are lemmas rather than word forms, the picture is not very different. The minimum number of instances needed for a rough outline of usage will rise to an average of about fifty for English (but many more for highly inflected languages).

If the research is about events which are more complicated than just word occurrence, then

the estimate of a suitable corpus size will also get more complicated. For example if the research is about multi-word phrases, it must be remembered that the occurrence of two or more words together is inherently far rarer than either on its own. So if each of the two words in a minimal phrase occur 20 times in a million word corpus, for 20 instances of the two together the arithmetic suggests a corpus of approximately 5 billion words will be needed. For three words together of this frequency the size of the corpus could be beyond our imaginings.

However, words do not occur according to the laws of chance, and if the phrases chosen are normal ones in the language, they will occur many times more often than the arithmetic projection above; so a much smaller corpus is likely to contain sufficient instances. To estimate roughly the size of a corpus for retrieval of a combination of two objects, first estimate the size you will need for the less common object on its own and then raise that figure by an order of magnitude. If there are 20 instances per million words for each of two words in a phrase, then twenty million words is likely to provide 20 instances of the pair (rather than the 5 billion projected by the arithmetic); if there are three of this frequency than 200 million words will probably be enough.

These are the kinds of figures that you will need to use in estimates of your optimal corpus size. Now we must build in the considerations of the methodology that you intend to use, because this can have a dramatic effect on the size.

The main methodological point is whether, having examined directly the initial results of corpus searches you intend to return to indirect methods and use the computer for further stages, recycling and refining early results[6]. If the latter, you will have to increase the minimum number of occurrences of your object quite substantially. This is because the regularities of occurrence that the machine will search for are not on the surface, and the way the computer works is to examine the *cotexts* minutely searching for frequently repeated patterns. Having found these it can then isolate instances of unusual and particular co-occurrences, which can either be discarded or studied separately after the main

patterns have been described. For example, if the computer searches for the adjectives that come between *in* and *trouble*, in text sequence (Bank of English 17/10/04) these are:

> *unspecified, terrible, deep, serious, deep, Cuba, serious, serious, great...*

It is reasonable already to anticipate that *deep* and *serious* are likely to be important recurrent collocates, but single instances of the others do not offer useful evidence. In fact *unspecified* does not recur, *terrible* is a good collocate, with 33 instances out of 1729. *Deep* is an important collocate with 251 instances, 14.5%, while *Cuba* is unique. *Serious* is slightly greater than *deep* at 271. *Great*, on the other hand, scores merely 8. The next in sequence is *big*, which at 235 instances is up with *deep* and *serious*. As we examine more and more instances, these three adjectives gradually separate themselves from all the others because of the number of times they appear — in total (757), almost half of all the instances. The nearest contender is *real*, at 142 quite considerably less common, and after that *financial* at 113. The computer also records as significant collocates *terrible* (35), *dire* (31) and *desperate* (28); *deeper* (14), *double* (14), *foul* (11), *bad* (14), *such* (28), *enough* (17) and *worse* (11).

The pure frequency picks out the three or four collocates that are closely associated with the phrase *in trouble*, and reference to the statistical test (here the t-score) adds another dozen or so adjectives which, while less common in the pattern are still significantly associated and add to the general gloom that surrounds the phrase. Single occurrences like *unspecified* and *Cuba* drop into obscurity, as do *terminal* (2) and *severe* (4), which occur among the first 30 instances.

The density of the patterns of collocation is one of the determinants of the optimal size of a corpus. Other factors include the range of ambiguity of a word chosen, and sometimes its distribution among the corpus components.

If you intend to continue examining the first results using the computer, you will probably need several hundred instances of the simplest objects, so that the programs can penetrate below the surface variation and

isolate the generalities. The more you can gather, the clearer and more accurate will be the picture that you get of the language.

8. Specialised corpora

The proportions suggested above relate to the characteristics of general reference corpora, and they do not necessarily hold good for other kinds of corpus. For example, it is reasonable to suppose that a corpus that is specialised within a certain subject area will have a greater concentration of vocabulary than a broad-ranging corpus, and that is certainly the case of a corpus of the English of Computing Science (James et al 1994). It is a million words in length, and some comparisons with a general corpus of the same length (the LOB corpus) are given in Table 1 (the corpus of English of Computing Science is designated as 'HK').

	LOB	HK	%
Number of different word-forms (types)	69990	27210	39%
Number that occur once only	36796	11430	31%
Number that occur twice only	9890	3837	39%
Twenty times or more	4750	3811	80%
200 times or more	471	687	(69%)

Table 1. Comparison of frequencies in a general and a specialised corpus.

The number of different word forms, which is a rough estimate of the size of the vocabulary, is far less in the specialised text than it is in the general one — less than 40% of its size. The proportion of single occurrences is another indication of the spread of the vocabulary, and here the proportional difference between the two corpora is even greater, with the specialised corpus having only 31% of the total of the other corpus. Word forms which occur twice are also much less common in the specialised corpus, but the gap closes quite dramatically when we look at the figures for twenty occurrences. At a frequency of 200 and above the proportions are the other way round, and the general corpus has only 69% of the number of such words in the specialised corpus. Assuming

that the distribution of the extremely common words is similar in the two corpora, these figures suggest that the specialised corpus highlights a small, probably technical vocabulary.

This is only one example, but it is good news for builders of specialised corpora, in that not only are they likely to contain fewer words in all, but it seems as if the characteristic vocabulary of the special area is prominently featured in the frequency lists, and therefore that a much smaller corpus will be needed for typical studies than is needed for a general view of the language.

9. Homogeneity

The underlying factor is homogeneity. Two general corpora may differ in their frequency profile if one is more homogenous than the other, while specialised corpora, by reducing the variables, offer a substantial gain in homogeneity.

Homogeneity is a useful practical notion in corpus building, but since it is superficially like a bundle of internal criteria we must tread very carefully to avoid the danger of vicious circles. As long as the choice of texts in a corpus still rests ultimately with the expertise and common sense of the linguist, it is appropriate for the linguist to use these skills to reject obviously odd or unusual texts. In any variety of a language there will be some texts — "rogue" texts — which stand out as radically different from the others in their putative category, and therefore unrepresentative of the variety on intuitive grounds. If they are included because of some high principle of objectivity, they are just wasted storage in the computer[7]. The principle of recurrence (see below) implies that a single occurrence of a feature is unlikely to be accepted as an authentic feature of a language or variety; hence unless texts share a large number of features the corpus will be of little use. There is a balance to be struck between coverage and homogeneity in the attempt to achieve representativeness.

The use of homogeneity as a criterion for acceptance of a text into a corpus is based certainly on the impression given by some features of its language, but is a long way from the use of internal criteria. A corpus

builder who feels that this criterion might threaten the status of the corpus can of course simply not make use of it, because it is really just a short cut. Rogue texts are usually easy to identify, and of course they must be genuinely exceptional; if we begin to perceive groups of them then it is our classification that must be re-examined, because we may have inadvertently collapsed two quite distinct text types.

It must be conceded at this point that we have moved in corpus design away from a completely objective stance and a blind reliance on objective external criteria. It is pleasant to envisage a utopia where corpora will be so large that a proportion of unique and strange texts can be included (if selected on objective criteria) without sabotaging the aims of the corpus design; if so this happy state of affairs is still quite a long way off. Such texts would largely disappear because their patterns would never be strong enough to be retrieved, so while corpora are still, in my opinion, very small indeed it is sensible in practical terms not to put "rogue texts" in at all. Provided that designers and builders accept the burden of documenting their decisions, there is little danger of distortion.

10. A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided.

10. Character of corpus research

It is necessary to say something here about the "typical studies" mentioned above, because at many points in this chapter there are assumptions made about the nature of the research enquiries that engage a corpus. This section is not intended in any way to limit or circumscribe any use of corpora in research, and we must expect fast development of new methodologies as corpora become more accessible and the software more flexible. But in any resource provision, the provider must have some idea of the use to which the resource will be put, and that is certainly so with corpora.

Corpus research is mainly centred on the recurrence of objects; initially surface entities

like word forms, objects can be re-defined after going through a process of generalisation, which means that forms which are not identical can be classified as instances of the same object. As noted above, the lemma is a clear example of this process.

Studies range from (a) close attention to textual interpretation, using only a few instances, through (b) the substantial quantities needed for language description on which the section above on "size" is based, to (c) large-scale statistical processing. All rely on recurrence as their starting point. The opposite of recurrence, uniqueness, cannot be observed with certainty in a corpus, because, as conceded near the beginning of this chapter, uniqueness in a corpus does not entail uniqueness in a language. However, very rare events can be, and are, studied, and of course the arithmetic of combinations means that most stretches of text that are more than a few words long are unlikely to recur, ever.

The use of a corpus adds quite literally another dimension to language research. If you examine a KWIC concordance, which is the standard format for reporting on recurrence, it is clear that the horizontal dimension is the textual one, which you read for understanding the progress of the text and the meaning it makes as a linear string, while the vertical dimension shows the similarities and differences between one line and the lines round about it. The main "added value" of a corpus is this vertical dimension, which allows a researcher to make generalities from the recurrences.

The current dilemma of much corpus linguistics is that the number of occurrences that a researcher can examine at once — in practice a screenful, some 23 lines — is a rather small amount of evidence, given the great variability of the language in use. On the other hand, to hand over the examination to the computer, where almost any number of instances could be processed very quickly, requires programming skills or a thorough knowledge of available software resources and how to make use of them. There is obvious advantage in getting the machine to do as much of the work as possible — in particular the gain in objectivity that results — but it requires much more investment in advance than the simple direct scrutiny of a

small sample.

11. What is not a corpus?

As we move towards a definition of a corpus, we remind ourselves of some of the things that a corpus might be confused with, because there are many collections of language text that are nothing like corpora.

The World Wide Web is not a corpus, because its dimensions are unknown and constantly changing, and because it has not been designed from a linguistic perspective. At present it is quite mysterious, because the search engines, through which the retrieval programs operate, are all different, none of them are comprehensive, and it is not at all clear what population is being sampled. Nevertheless, the WWW is a remarkable new resource for any worker in language (see Appendix), and we will come to understand how to make best use of it.

An archive is not a corpus. Here the main difference is the reason for gathering the texts, which leads to quite different priorities in the gathering of information about the individual texts.

A collection of citations is not a corpus. A citation is a short quotation which contains a word or phrase that is the reason for its selection. Hence it is obviously the result of applying internal criteria. Citations also because lack the textual continuity and anonymity that characterise instances taken from a corpus; the precise location of a quotation is not important information for a corpus researcher.

A collection of quotations is not a corpus for much the same reasons as a collection of citations; a quotation is a short selection from a text, chosen on internal criteria and chosen by human beings and not machines.

These last two collections correspond more closely to a concordance than a corpus. A concordance also consists of short extracts from a corpus, but the extracts are chosen by a computer program, and are not subject to human intervention in the first instance. Also the constituents of a corpus are known, and searches are comprehensive and unbiased. Some collections of citations or quotations may share some or all of these criteria, but

there is no requirement for them to adopt such constraints. A corpus researcher has no choice, because he or she is committed to acquire information by indirectly searching the corpus, large or small.

A text is not a corpus. The main difference (Tognini Bonelli 2001 p.3) is the dimensional one explained above. Considering a short stretch of language as part of a text is to examine its particular contribution to the meaning of the text, including its position in the text and the details of meaning that come from this unique event. If the same stretch of language is considered as part of a corpus, the focus is on its contribution to the generalisations that illuminate the nature and structure of the language as a whole, far removed from the individuality of utterance.

12. Definition

After this discussion we can make a reasonable short definition of a corpus. I use the neutral word "pieces" because some corpora still use sample methods rather than gather complete texts or transcripts of complete speech events. "Represent" is used boldly but qualified. The primary purpose of corpora is stressed so that they are not confused with other collections of language.

A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.

Acknowledgements

This chapter relates historically to a paper entitled *Corpus Creation* which was presented to the Council of Europe in February 1987; it was revised for publication in Sinclair (1989) and updated again as a chapter in Sinclair (1991). After a further decade it has been completely rewritten, but covers much the same ground as the earlier papers.

I am grateful to Knut Hofland, Sattar Izwaini and Martin Wynne for help with Table 1 and the other statistics.

Notes

1. See the brief discussion on homogeneity later in this chapter.

2. For a discussion of the role and limitations of intuition, see Sinclair (2004).

3. See the MICASE website, http://www.hti.umich.edu/m/micase/ under "Speech event and speaker categories", which is a very elaborate classification, leading to small cells and many empty ones.

4. For further discussion of this point see Sinclair 2004.

5. Knut Hofland reports that in the LOB corpus there are 25,992 instances of tag/word combinations that occur once only, as compared with 36,796 word forms (see Table 1). While the lemmatisation reduces the number of different objects, the tag assignment increases the number by giving more than one tag to the same form.

6. There is a brief treatment of this point in Sinclair (2001).

7. One good example of a rogue text appeared in one of the earliest specialised corpora — Roe's corpus of textbooks in physical science (Roe 1977). This million-word corpus held a dozen or so full-text documents, which showed considerable homogeneity, all except one. The rogue text turned out to be an Open University textbook, and it reflected the innovative style of OU teaching, the new student body attracted to the OU, and the resolve to make learning a part of life. So any generalisation that applied to all the other texts was not supported by the OU text, and virtually none of its features were found in the others. The contrast was so marked that Roe had to make a separate statement for the OU text and one for all the rest. The text excluded itself, which was ironic because Roe had chosen it deliberately as an example of good current communication; its "rogue" status has nothing to do with its worth as an academic textbook, but shows the sharp difference in approach that is associated with the OU.

Continue to Chapter Two: Adding Linguistic Annotation

Continue to John Sinclair's Appendix: How to build a corpus

Return to the table of contents