Table of Contents

Previous Chapter

Bibliography

Sections in this chapter:

Developing Linguistic Corpora: a Guide to Good Practice

Appendix: How to build a corpus

John Sinclair, Tuscan Word Centre

Introduction

The job of corpus building divides itself into two stages, design and implementation, but these cannot be completely separated, for reasons which are largely practical.

One is the cost. Nowadays most corpora are put together from text that is already digitised; the cost of putting into electronic form text which only exists on paper is very much greater than the cost of merely copying, downloading and gathering data that is already digitised; so there has to be a compelling reason for using any of the more laborious methods which were used to capture data in the days before electronic text.

Sometimes, however, it is necessary, to do things the hard way; for a corpus of informal conversations, for example, or historical documents or handwritten or manuscript material. But in all such cases it is worth a serious search of various collections and archives, and perhaps a query on the professional lists, before undertaking the labour of entering new text.

Another reason for mixing principle and practice in corpus building is because some kinds of data are inherently difficult or even impossible to obtain, and a measure of compromise is often necessary; some authors categorically refuse to have their work stored in a corpus or insist on high fees; some types of interaction are extremely difficult to make records of; in many countries surreptitious recording is illegal;[1]some documents that use graphics are unscannable and have to be unpacked before being laboriously typed in to the corpus.

For languages that are used in substantial segments of the globe there will be found a very large amount of text material on the internet. Even for smaller languages there is often a remarkable amount and range of material. If the electronic resources available are not adequate then the least expensive alternative is scanning printed texts; however this is time-consuming and the output from the scanner needs to be edited at least superficially. See below on Perfectionism.

The worst option is to have to type in large amounts of textual material; this is still unavoidable with transcripts of spoken interaction, but requires a consumption of resources that drags a project, limits its size and reduces its importance. Keying in may be a viable option for individual texts which are not available in digital form and which are not easy to scan, but for a large text corpus, there are likely to be easier options.

The World Wide Web

While web pages are likely to be the most immediately accessible sources of material, they are by no means the only source, and some of the most valuable text material is merely indexed on a web page, requiring further searching. For example, many large document archives put up their catalogue on the web, and give opportunities for downloading in various formats. Here the web is playing the role of a portal. Other providers of text data may issue CDs, especially when there is a lot of data to be transferred. Sometimes payment is required, especially for material that is popular and under copyright; corpus builders should consider carefully the costs of such data and whether it is justified.

Also available on the internet are many — probably millions — of documents that are circulated by e-mail, either messages or attachments. By subscribing to appropriate lists, your collection of material can grow quickly.

The Web is truly bountiful, but it is important to appreciate that the idea of a corpus is much older than the Web, and it is based on "hard-copy" concepts, rather than cyber-objects like web "pages". A corpus expects documents (including transcripts) to be discrete, text to be linear and separable from non-text, and it expects documents to fall into recognisable sizings, similar to hard-copy documents. A normal corpus has no provision for hypertext, far less flashing text and animations. Hence all these familiar features of the Web are lost unless special provision is made to retain them. The procedural point (1) — see below — is relevant here; the documents in their original format should be carefully preserved; it is up to the corpus managers how far hypertext links are preserved as well in a "family" of documents, but, like all the other texts in a corpus, the Web document is ultimately removed from the environment of its natural occurrence.

Some projects are learning how to make multimedia archives within which spoken or written text is one of the data streams, and a more modern notion of a corpus may result from this research. Linguists need make no apology, however, for concentrating on the stream of speech or the alphanumeric stream; particularly in the early stages of a new discipline like corpus linguistics the multimedia environment can be so rich that it causes endless diversions, and the linguistic communications can get submerged.

At present it is important to know precisely what is actually copied or downloaded from a web page. This is not always obvious, and quite often it is not at all the document that is required. The "source" file, which contains all the mark-up, is easy to download but difficult to handle; "text-only" or "print-friendly" versions of a page can be helpful. In all cases it is essential to review what you have tried to capture to make sure that it is the target document — it may only be the address, or a message such as "page not found".

The cheerful anarchy of the Web thus places a burden of care on a user, and slows down the process of corpus building. The organisation and discipline has to be put in by the corpus builder. After initial trials it is a good idea to decide on a policy of acquisition and then stick to it as long as it remains practical; consistency is a great virtue as a corpus gets larger, and users of a corpus assume that there is a consistency of selection, processing and management of the texts in the corpus. Already we read a lot of apologies for the inadequacies of corpora, often inadequacies that could have been avoided.

Another tricky question is that of copyright — not the familiar copyright of publications, but the more nebulous issue of electronic copyright. In principle, under UK law, publication on the internet confers the rights on the author whether or not there is an explicit copyright statement. Every viewing of a web page on a screen includes an act of copying. If there is doubt, contacting the named copyright holder is advisable.

Perfectionism

So when you design a corpus it is probably best to write down what you would ideally like to have, in terms of the amount and the type of language, and then see what you can get; adjust your parameters as you go along, keeping a careful record of what is in the corpus, so that you can add and amend later, and if others use the corpus they know what is in it.

It is important to avoid perfectionism in corpus building. It is an inexact science, and no-one knows what an ideal corpus would be like. With good research on such matters as the penetration of documents in a community, our present guesswork can certainly be improved on, and even the influence of the spoken word relative to the written word may be estimated more securely than at present. Until then compilers make the best corpus they can in the circumstances, and their proper stance is to be detailed and honest about the contents. From their description of the corpus, the research community can judge how far to trust their results, and future users of the same corpus can estimate its reliability for their purposes.

We should avoid claims of scientific coverage of a population, of arithmetically reliable sampling, of methods that guarantee a representative corpus. The art or science of corpus building is just not at that stage yet, and young researchers are being encouraged to ask questions of corpora which are much too sophisticated for the data to support. "It is better to be approximately right, than to be precisely wrong."[2]

We should also keep a distance from claims of accuracy of analysis by current software. Even what seems to be almost perfect accuracy is likely to be systematically inaccurate, in that whole classes of data are always misclassified. This problem arises because accuracy is in the mind of the analyst, and may not correspond with the distribution of patterns in the corpus. Furthermore, in a corpus of, say, a hundred million words, 99% accuracy means that there are more than a million errors.

Indicative, not definitive

The results of corpus research so far are indicative of patterns and trends, of core structures and likely contributions to theory and description, but they are not yet definitive. It should become a major objective of serious corpus research to improve the procedures and criteria so that the reliability of the descriptive statements increases. However, this move to greater maturity of the discipline is not an admission of any limitations; we established above that corpora are ultimately finite and that this is a positive property of them, giving them descriptive clarity. A description based on an adequate theory and a very large and carefully built corpus, combined with flexible and theory-driven software will provide descriptions far above what we live with at present. The fact that, like any other conceivable description, they will not reflect the ultimate flexibility and creativity of language will be of interest to a small group of specialists, no doubt, but not to the mainstream of research.

Corpus-building software

If you ask Google for "corpus builder" — at the time of writing — you get a number of useful leads which can support corpus-building activities, for example, recognising a particular language in order to select only texts in that language. Software is also offered which will build a corpus for you, index it and allow searches of various kinds. More and more of this kind of product is likely to appear, and according to your purposes and resources you may look into suitable packages. Some are commercial ventures and sell at up to a thousand euros or so, and these normally allow a trial period, which is worth investigating. Study the small print carefully, looking for limitations of size, speed and flexibility, and make sure that the software will perform as you want it to.

Free or open-source software is often more specialised than the commercial products, but is more likely to be tricky to install and is not always friendly to use, so be prepared for some initial problems with this.

Procedure

The main considerations that affect the choice of texts in a corpus are given in the paper above under "Representativeness". Once a text has been chosen for a corpus, and the location of a copy in some usable format has been determined, then there are several recommended steps towards making a useful and understandable corpus.

1. First, make a **security copy** of the text, in exactly the format received. If it is not in electronic form, keep the hard copy version for later reference[3],
2. Save the text in **plain text** format. Sometimes this is not straightforward, and in extreme cases a text may have to be rejected if its formatting cannot be standardised. But most text packages have a plain text option. This step is recommended even if your intended processing package will handle a mark-up language like HTML, XML, SGML, or word processor output. The issue is flexibility — conventions keep changing, new ideas come in and suddenly everything is old-fashioned. A corpus consisting of texts in a mixture of formats is impossible to handle. Conversion from one format to another is usually laborious and

uncertain, no matter what the optimists say. Plain text, the rock-bottom linear sequence of letters, numbers and punctuation marks, is almost always an easy conversion, and that is the one to keep.

3. Provide an **identification** of the text at the beginning of it. The simplest identification, and the one that makes the least disruption to the text, is a short reference — just a serial number, for example — to an off-line database where relevant information about the text is stored. More elaborate identifications are called headers, and these can be elaborate structures where information about author, date, provenance etc. is added to the text. To keep the added material separate from the text material, a corpus with headers has to be coded in a mark-up format, such as one of those mentioned above. The mark-up allows the headers and other additions to be ignored when the corpus is searched.
4. Carry out any **pre-processing** of the text that is required by the search software. The proprietary software intended for small corpora on a Windows platform normally includes all necessary steps in processing. For large corpora using Linux-type platforms, search engines frequently specify some initial processing to make the most popular retrieval tasks quick and efficient, e.g. the compilation of a list of all the different word-forms in the text, to be used as a basis for wordlists, concordances and collocational profiles.
5. When the corpus is complete, at least so that you can get started with research on it, make a security copy on a CD; as you add to an initial corpus, make further CD copies so that you can always restore the corpus following a disk crash (see also the advice in Chapter 6 on archiving and preservation). Check your working version from time to time because mysterious corruption can affect your files. Always check the integrity of the corpus if it gives you strange results.

Notes

[1]. In those countries that tolerate surreptitious recording, there is still the ethical issue of privacy, and anyone handling data of this kind should consider (a) offering the participants the option of deleting the recording after the event, (b) physically removing from the tape any passages which could be used to identify the participants.

[2]. This is Rule 8 (Use Common Sense) of the 9 Rules of Risk Management published as an advertisement by the RiskMetrics Group, cited in *The Economist* of April 17th 2004, page 29.

[3]. It is well within present technology practice to make a facsimile of a printed page and to align it with an electronic version of the text that is printed on it; the user could then call up the physical image of the page at any time to resolve any issues of interpretation. This would do away with the need to have formatting mark-up tags in the text stream, at least in cases where the text is derived from a printed original.

Continue to Chapter Two: Adding Linguistic Annotation

Continue to Bibliography

Return to the table of contents