Corpus building and investigation for the Humanities:

An on-line information pack about corpus investigation techniques for the Humanities

Unit 2: Compiling a corpus

David Evans, University of Nottingham

2.1 Designing a corpus

If you have decided that using a corpus will help you with your research but that no corpus already exists which is suitable for your purposes, you will need to design your own corpus. In order to build a corpus there are a number of factors which need to be taken into consideration. These include size, balance and representativeness and will be discussed below.

Size: The size of the corpus depends very much on the type of questions that are going to be asked of it. As a rule of thumb, bigger is generally considered to be better as the software can be instructed to filter out some of the output. However, it is possible to get much useful data from a small corpus, particularly when investigating high frequency items. In fact this may be desirable to do this rather than being overwhelmed by too much data from a big corpus. It is also worth bearing in mind that some corpus software set limits on the number of concordance output lines; when they get to these limits they simply stop searching the corpus.

You may also be constrained by more practical considerations. If you need to transcribe spoken data with a high degree of detail then it may only be feasible to work with thousands rather than millions of words. With written texts you may be limited by what you can obtain permission for from the copyright holder.

Balance: Let's imagine that you wanted to build a corpus of British broadsheet newspapers over a given time period. Your first instinct might be to simply load up all the articles from the newspapers into your corpus and get started. However, doing this raises questions of balance. Articles from one section of the newspaper may be much longer than those in another and may lead you to conclude that a particular feature is much more common in a particular type of journalism, where proportionally this is not actually the case. Instead you may wish to sample texts up to a predetermined word limit. Research has suggested that samples of 2,000 to 5,000 words are sufficient. However, it is also worth bearing in mind that introductions, bodies and conclusions differ linguistically and so samples would need to be balanced for this too.

If you are collecting data for a spoken corpus, it may be necessary to carefully consider the types of people you use as informants. This will allow you to decide if your data balanced in terms of the sex, class, age, ethnic background, etc. of the participants and thus how representative any claims you might make will be of the wider population. Getting this balance right is not an exact science and there are no reliable ways of determining whether a corpus is truly balanced. One approach to achieving balance is to use an existing corpus as a model. Unit 3 has a number of hyperlinks to corpus websites where this information is readily available.

Representativeness: A corpus can be said to be representative if the findings from that corpus are generalisable to language or a particular aspect of language as a whole. Obviously, it is not possible to collect an entire language to test the representativeness of a corpus. Instead we can use the notion of 'saturation' (also known as 'closure', see McEnery et al 2006:15-16). Saturation (at the lexical level) can be tested for by taking a corpus and dividing it into equal sections in terms of number of words. If another section of the same size is now added, the number of new items in the new section should be approximately the same as in the other sections.

In general it is important to be pragmatic throughout the construction process as there may well be unexpected problems along the way. The process of building a corpus is a cyclical one. As you learn more apply this knowledge to the whole corpus and be prepared to make changes, including leaving out data you have gathered, if this improves the final corpus. Keep a detailed record of the data you collect. If you are collecting spoken data make a note of the speakers' names, sex, age, ethnic background, etc. This information may not seem useful initially but it can be incorporated into the corpus itself at a later date and used for a wider range of research.

2.2 Why you need your texts in electronic form

In the previous unit we established that the texts in a corpus need to be in electronic form. In this section we will discuss the parameters of that form in a little more detail. Most corpus investigation software will not read the kind of complex embedded formatting associated with common word processing packages like Microsoft Word or pdfs. Whilst the technology is constantly changing, at present most corpus documents are saved in .txt format.

It is important to remember that any document that is prepared for corpus analysis is only a representation of the original. Because of this much of the contextual information that producers and receivers of both spoken and written texts take for granted may well be lost when these texts are rendered in a corpus-friendly form. The usual way to deal with this problem is to add some annotation.

Annotation refers to all the extra information that is added to the texts in order to aid the researcher to retrieve as much relevant information as possible. When considering annotation we can divide a corpus text into two sections: the header and the body.

The header often contains metadata – that is things like the name of the author, the title of the work, the year of publication, etc., in the case of written texts, or the age, sex, ethnicity, social background or role of the speakers in a spoken context. The inclusion of this information makes the corpus a much more powerful research tool as it means elements of the language in the corpus can be investigated not just as purely linguistic but also as social phenomena.

Within the text itself it is also possible to add annotation, known as tagging where each word in the corpus is tagged with a code which gives the user extra information. Many commercially available corpora are tagged for part of speech (POS), which is done by a combination of automated and manual means. Other corpora have been annotated to provide extra semantic, stylistic or pragmatic information or tagged for error in the case of learner corpora. Such annotation can be viewed or blocked by the software depending on the user's preference.

When viewing text via corpus software in the form of concordance lines, it is not usually possible to see textual features such as sentence and paragraph boundaries. These can be added back in through annotation in a form like this:

<s>The cat sat on the mat<\s>

This annotation also allows the researcher to locate sections of the text quickly and find out what linguistic features are present at the beginnings and endings of sentences or paragraphs.

One issue to consider when using characters other than standard alphabet is whether these characters will be recognised by your corpus browsing software and, if not, how they might be represented in the text. If you are likely to be doing a substantial amount of work with non-alphabet characters then it would be worthwhile looking at software that makes use of Unicode.

2.3 Issues in getting texts from the Internet

Probably the easiest way of obtaining texts already in electronic format is to download them from the internet. Most web pages can be easily rendered as a text file, using either the computer's clipboard or the Save As function. This will usually strip out much of the unwanted formatting but may also leave a significant amount of superfluous textual information such as menus, links or large areas of blank space and will require further, manual editing.

This approach is useful if the intention is to collect a small amount of specific data. However, if you are planning to build a corpus using a large number of web pages then it would be preferable to automate this process. It is possible to capture entire websites, rather than just individual pages, using an offline browser. There are a wide variety of offline browsers available, many as freeware, which can not only download whole sites, but also archive related web pages based on topic-specific searches.

WordSmith Tools 4.0 has a WebGetter function which works with a user-specified search engine to trawl the internet for a search term and then download the 'first 100 sources or so' (Scott 2006: online). It also has a minimum word count function which discards pages which do not have enough text to be useful for corpus compilation. There are two drawbacks with this system: Just like any normal search engine search, there is no way to guarantee the quality of the hits that are found. Secondly, and this may also apply to offline browsers, the pages are saved as HTML and therefore need to be 'cleaned up' for use in your corpus.

To convert HTML into a useful format (i.e. text files) you may want to use the Multilingual Corpus Toolkit (MLCT) (see McEnery et al 2006: 74). This is available free from http://www.routledge.com/textbooks/0415286239/Zip/MLCT.zip and requires Java 1.4 or above. It has several additional features, including part of speech tagging which may be useful for the corpus builder.

Baker (2006: 36-7) points that CMC (computer mediated communication) is attracting increasing academic interest. However, even though a text may have been created in electronic form, there are caveats that the corpus builder must take note of. Instant messaging uses a large number of 'smilies', which are an important part of the message. Users of discussion groups may have an avatar in the form of an image, which are difficult to translate into text but is often an important statement of that user's identity. Email responses often retain segments of the message which is being replied to or make reference to attached text documents, photographs or soundfiles. Decisions will have to be taken on how much of this can or should be retained in a corpus. The increasingly multi-modal nature of the Internet poses many interesting challenges for the corpus builder.

2.4 Issues in scanning and keying in texts

You may wish to compile a corpus of data that does not already exist or is not readily available in electronic form. There are two ways in which you can transfer such texts into electronic documents, by scanning them using OCR (optical character recognition) software or by keying them in manually. Both methods have their advantages and their downsides.

In the late 1970's when researchers at the University of Birmingham were working on what became the Bank of English, they purchased an early scanner at a cost of around £70,000. Thankfully, nowadays scanners with OCR are easily available, increasingly widespread and can cost as little as £30. Scanners can save even the best typist a great deal of time and effort, especially where large quantities of text are concerned. However, as anyone who has used OCR will testify, it is still too inconsistent to be fully relied upon. All scanned texts must be carefully proofread before they can be used as part of your corpus.

Another area to consider before scanning a document is the way the page is laid out; whether the text uses bold type and italics or different type faces. Also, are there any non-textual elements such as pictures, diagrams or graphic data? The output from scanning a document is often in .doc or .pdf form and it may attempt to replicate the original, using complex

formatting to do so. This formatting will not transfer over to a text file so you will need to consider whether or not you need to represent it in some other way or simply cut it.

If the source texts are handwritten documents such as personal letters then the only way to get these into electronic form may be to key them in. This option is usually thought of as a last resort (Baker 2006: 35) and it may be necessary to consult the phone book or the internet for local professional typing services if your own skills are not up to the job.

One thing that must be considered when using any written text as part of a corpus is copyright. Anyone intending to put together a corpus for commercial purposes must always obtain the permission from the publishers of the source texts. Many commercially available corpora contain texts from a large number of sources and obtaining permission to use these can be a very long-winded and financially costly process. However, if you are building your own small-scale corpus there a few things you can do to make getting permission a little easier. Point out to the copyright holder that:

- Texts will only be used for non-profit research purposes.
- You are the only person who will have access to the data.
- If necessary, the texts will be deleted after a certain period should this make permission easier to obtain.

Also, remember that:

- It is much easier to obtain permission if you are only using a part of a text.
- If the project you are working on is financed by a funding body, you should also check with them whether they have their own rules regarding permissions.

2.5 Issues in compiling a spoken corpus

The availability of recording equipment has had a major impact on the study of spoken discourse. Continued advancements in digitalisation and miniaturisation of recording technology mean it is now relatively easy and affordable to collect spoken data. And whilst building a corpus of data that you have personally collected and transcribed can be a very rewarding, it is also an extremely time-consuming process and is usually not the kind of project that can be undertaken without a lot of planning and execution time.

Unlike writing, the nature of spoken discourse means that it is subject to the observer's paradox. If the goal of your research is to collect what might be described as 'natural' or 'real' data, there are a number of issues to consider. Obviously, the best way to get this kind of data is for the participants in your study to be unaware that they are being recorded. Surreptitious recording has been used by corpus linguists in the past but is now regarded as unethical at best and, in some circumstances, may well be illegal.

Instead the compilers of many recent corpora have asked contributors to wear lapel microphones and carry recorders around with them for a part of their day. The data captured using this method is often characterised by questions regarding the microphone from other interlocutors at the outset of conversations, but these do not last long and conversations tend to proceed as normal.

Before undertaking this kind of recording you will need to check that the recording equipment you have access to is compatible with the environments you will be recording in. Excessive background noise that the human ear would filter out might overpower voices on a recording and the data would be lost. Test recordings are always advisable.

Whilst casual conversation is not protected by copyright, you should always make sure you obtain written permission from anyone who agrees to take part in your recordings. Make sure, as far as possible, the participants know what your research is about, who will have access to their contributions and whether you intend to anonymise the transcripts. For further advice consult the BAAL ethics guide on their website.

When transcribing spoken data it is very important to decide on the level of complexity that will be required of the data from the outset. At its simplest a transcription can simply be an

orthographic representation of the words uttered by the informant and would thus probably resemble a play script. However, it may be necessary to transcribe more of the features that are common in unscripted speech such as pauses and hesitations, false starts, repetition and passages of inaudible speech. Many professional corpus compilers have made their transcription conventions available which you are free to use or adapt.

Finally, check your transcription for consistency, especially if you have employed someone else to do the transcribing.

References and further reading:

Baker, P. (2006) *Using Corpora in Discourse Analysis* London: Continuum

Biber, D., Conrad, S. & Reppen, R. (1998) *Corpus Linguistics* Cambridge: Cambridge University Press

Leech, G., Myers, G., Thomas, J., (1995) *Spoken English on Computers* Harlow: Longman

McEnery, T., Xiao, R. & Tono, Y. (2006) *Corpus-Based Language Studies* Abingdon: Routledge

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

Scott, M. (2006) WordSmith Tools [available at http://lexically.net/wordsmith/index.html]

Thompson, P. (2004) Spoken Language Corpora in Wynne, M. (ed.) *Developing Linguistic Corpora: a Guide to Good Practice* [available at http://ahds.ac.uk/creating/guides/linguistic-corpora/index.htm]

BAAL ethical conventions
http://www.baal.org.uk/about_goodpractice_full.pdf
http://www.baal.org.uk/about_goodpractice_stud.pdf

BASE Corpus transcription conventions:
http://www2.warwick.ac.uk/fac/soc/celte/research/base/conventions.doc