CORPORA AND CONCORDANCES: COLLECTING LINGUISTIC DATA FOR ESSAYS

Concordances have been used for hundreds of years for studying the use of words and phrases. Until the invention of computers and associated software, concordances were prepared by hand. This was clearly very time-consuming and done only for texts of great cultural value, usually major religious and literary texts, such as the Bible and Shakespeare.

CORPORA AND CONCORDANCE SOFTWARE ON THE WEB

Nowadays, computers can be used. Large collections of contemporary texts and search programs are available, and concordances of words and phrases can often be prepared in a few minutes. Some large corpora and concordance software are available on the world-wide-web, and allow you to search millions of words of spoken and/or written English. This is very useful for collecting examples for linguistics essays ... ;-)

The following address gives access to the **British National Corpus** (BNC), a 100-million word corpus of British English (90 million words written and 10 million words of spoken). It allows you to find up to 50 examples of words and phrases, and print them out, or save them to disk.

http://www.natcorp.ox.ac.uk/

The following two alternative addresses are for an interactive data-base, **Phrases in English** (PIE), which allows you to study frequent phrases in the BNC. The data-base contains all phrases, up to 8 words in length, which occur more than 3 times in 100 million words. It allows searches for many kinds of patterns. There is a detailed introduction and tutorial.

http://pie.usna.edu
http://www.phrasesinenglish.org/

The following address gives access to the **CobuildDirect** corpus, a 56-million word corpus of spoken and written British and American English. Cobuild stands for "Collins Birmingham University International Language Database". It allows you to find up to 40 examples of words and phrases (in each of different sub-corpora), and print them out, or save them to disk.

http://www.collins.co.uk/Corpus/CorpusSearch.aspx

Even corpora of millions of words may contain only a few examples of rare words and phrases, and very large text collections may then be necessary.

There are also concordancers which search for words and phrases in the millions of pages in the world-wide-web, and then concordance these. One such concordancer is available at

http://www.webcorp.org.uk/wcadvanced.html

ACKNOWLEDGING YOUR SOURCES

If you do use such data in essays, you must acknowledge the source of your data.

First, you must acknowledge the work of those who have prepared the corpora, (just as you acknowledge the authors of articles and books).

Second, you must say what corpus you have used, so that the reader knows the contents of the data: whether the data are spoken or written, representative of British or American English, etc.

A BRIEF EXAMPLE

A concordance program searches for a word (or phrase or grammatical construction) and lists all examples along with the lines of co-text which it occurs in. The lines can be ordered alphabetically or in other ways. Here is a short fragment of a concordance of the word-form *undergo*.

```
01  women would have had to undergo a deep and important change o
02 ld people were likely to undergo a major psychological upheava
03 ing, had been induced to undergo a medical examination to see
04 k, each operative had to undergo a stringent medical examinati
05 er of the shop seemed to undergo a transformation. The rush wa
06  Mr Forbes was forced to undergo an emergency operation to rem
07 take kindly to having to undergo an identity check before bein
08 ly anyone at risk should undergo confidential testing on a tra
09 RAF widows would have to undergo 'demeaning means tests' years
10 ital and insisted that I undergo extensive tests. There was he
11 cers and men have had to undergo great privations. They landed
12  with two recessions and undergo immense change in that proces
13  Many of these creatures undergo intolerably cruel conditions
14 te employees may have to undergo lie detector tests. Rapist w
15 ctured skull may have to undergo neuro-surgery if his conditio
16 eans he will not have to undergo the punishing marathon of the
17 ind themselves having to undergo the painful dislocation entai
18 f they were expecting to undergo surgery, or if they had a his
19  they would also need to undergo years of specialized training
```

Note how *undergo* occurs

- to the left, with phrases such as *had to* and *forced to*
- to the right, with words for unpleasant things such as *tests* and *operations*.

These data are a small illustrative sample of concordance lines from CobuildDirect and the BNC.

Note also: I have numbered the lines and alphabetized them to the right. If you are presenting such concordance data for a seminar or a term paper, please number and order the lines in some appropate way. Otherwise it is very difficult to refer to individual lines of data.

CONCORDANCE SOFTWARE

You can also concordance your own files directly by using many different concordance programs. One entirely free concordancer is available from Laurence Anthony. It will make concordances and word-lists of input files (and also identify n-grams, frequent collocates, plot the distribution of words in texts, etc). It does quite a lot of the things which WordSmith Tools does, but does not cost anything ... You can download it from

http://www.antlab.sci.waseda.ac.jp/software.html

Or you can find the page it by googling "AntConc". This leads you to a download site for an *.exe file. This just has to be copied to your hard disk.

You could, for example, use the BNC site (above) to get 50 examples of some word or phrase which you are interested in, store the results, and then read them into AntConc.

NOTE ON FREE TEXTS OF WORKS OF ENGLISH LITERATURE

If works are out of copyright (i.e. basically pre-1900), they can often be obtained in a computer-readable, plain text format from

• Project Gutenberg:	http://promo.net/pg/
• Bartleby:	http://www.bartleby.com

You can download texts from these two sites and run the concordance software directly on these other texts. Texts for concordance programs must normally be in plain text format.

REFERENCES

The following books discuss corpora and concordances, and give many examples of analyses.

Barnbrook, G (1996) *Language and Computers*. Edinburgh UP.
Biber, D et al (1998) *Corpus Linguistics*. CUP.
Kennedy, G (1998) *An Introduction to Corpus Linguistics*. Longman.
Partington, A (1998) *Patterns and Meaning: Using Corpora for English Language Research and Teaching*. Benjamins.
Sinclair, J (1991) *Corpus Concordance Collocation*. OUP.
Sinclair, J. (2004) *Trust the Text*. London: Routledge.
Sinclair, J. ed (2004) *How to Use Corpora in Language Teaching*. Benjamins.
Stubbs, M (1996) *Text and Corpus Analysis*. Blackwell.
Stubbs, M (2001) *Words and Phrases: Corpus Studies of Lexical Semantics*. Blackwell.

Michael Stubbs
FB2 Anglistik, University of Trier, D-54286 Trier, Germany

This file last up-dated February 2008.
