Extracting Collocations from Specialised Corpora

Michael Barlow[1] and Ute Römer[2]

Abstract

As the growing number of publications (see Evert 2005) on the topic indicates, collocations, which can be defined as frequently occurring contiguous or non-contiguous combinations of words, are of central interest in linguistic analysis and description. Even though corpora and concordance packages offer important insights into the co-selectional tendencies of words, there is as yet no definitive way to locate collocations in a corpus.

However, there is now becoming available a new generation of software tools that enable users to extract from a corpus lists of candidate collocations for inspection. One of these new-generation tools is *Collocate* (Barlow 2004). *Collocate* uses frequency information and statistical analyses (t-score, log likelihood, MI) in order to retrieve lists of:

(a) collocations with a specified search word and within a set span (e.g. four words),
(b) n-grams (lexical bundles) of different lengths, and
(c) collocations extracted from the corpus as a whole.

In this workshop, we will first demonstrate some of the *Collocate* facilities, focussing on the extraction of terminology and meaningful items from specialised corpora, e.g. MICASE (the Michigan Corpus of Academic Spoken English). We will show in what ways *Collocate* can be used to provide insights into the special characteristics of different text types, and how the programme can highlight which word combinations members of a particular discourse community tend to use in order to create particular meanings.

In the hands-on part of the workshop, the participants will be able to work with and evaluate different functions and statistics used in *Collocate* in order to produce and interpret lists of collocations from different kinds of specialised corpora. (Workshop participants are also welcome to bring their own corpus data for use with *Collocate*.)

References

Barlow, M. 2004. *Collocate 1.0: Locating collocations and terminology.* Houston, TX: Athelstan.

Evert, S. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* Stuttgart: Institut für maschinelle Sprachverarbeitung. Available fromhttp://elib.uni-stuttgart.de/opus/volltexte/2005/2371/; accessed 14 January 2007.

[1] University of Auckland, New Zealand

[2] University of Hanover, Germany, *e-mail*: ute.roemer@engsem.uni-hannover.de