(Association) measure for measure: Evaluating the corpus-based identification of lexical collocations

Lexical collocations – in the Firthian sense of recurrent habitual word combinations – are a complex phenomenon for which neiter a satisfactory linguistic definition nor a convincing lexicographic operationalisation has been found yet. However, collocations are pervasive in language and therefore important for advanced language learners who wish to sound fluent and natural. Specialized collocation dictionaries such as the BBI Combinatory Dictionary of English (BBI) and the Oxford Collocations Dictionary for students of English (OCD2) have been produced to fill this need.

For these reasons, the automatic identification of lexical collocations from electronic corpora is of great interest to linguistic theory, practical lexicography and natural language processing. One of the most useful indicators of collocability is the tendency of two words to co-occur in text, which can be quantified with the help of a range of statistical association measures. Other indicators that have been used for the extraction of lexicalized multiword expressions – such as non-compositionality and non-modifiability – are not directly aplicable to Firthian collocations (because of their lack of clear-cut defining characteristics).

In my talk, I present the results of a recent evaluation study focused on lexical collocations, using the BBI and OCD2 collocation dictionaries as a gold standard based on the intuitions of professional lexicographers. In contrast to previous studies, this evaluation includes a systematic comparison of different source corpora (ranging from small and clean reference corpora to huge and messy n-gram databases), different collocational spans (from 1 to 10 words) and different levels of automatic annotation (dependency-parsed vs. part-of-speech tagged). One particular focus is the question to what extent Web corpora (such as ENCOW) can substitute for and improve on traditional reference corpora like the BNC, and whether "bigger is better" in quantitative corpus linguistics.