Uncovering the Extent of the Idiom Principle

Chris Greaves[1] and Martin Warren[1]

Abstract

One of the most important findings, if not the most important finding, to come out of corpus linguistics has been what Sinclair (1987) terms 'the idiom principle', i.e. the phraseological tendency, whereby words are co-selected by speakers and writers which gives rise to collocation and other features of idiomaticity. This paper describes and defines a new way of identifying and categorising word associations, the concgram, which is all of the permutations of constituency variation and positional variation generated by the association of two or more words. Concgrams therefore ignore traditonal distinctions between lexis and grammar. Using specially designed software (Concgram©, Greaves 2005), the concgrams of a corpus are identified and generated automatically without prior input from the user, other than to set the size of the span. Studying concgram search results reveals word associations in a way that other searches do not. In the case of the latter, attention is primarily drawn to the user- nominated node word(s), a popular and traditional starting point for corpus queries which is replaced by the notion of 'origin' in concgram searches where the focus of attention is on word associations and their constituency and positional variations.

Preliminary searches on a one-million-word corpus of spoken discourses have found that the majority of concgrams are made up of non-contiguous collocations, and show both constituency (AB, ACB) and positional (AB, BA) variations which can be calculated and sorted by frequency (Cheng, Greaves and Warren, 2006). Although contiguous collocations are also found in concgram searches, since many collocational patterns never occur contiguously, searches which focus on contiguous collocations present an incomplete picture of the word associations that exist. Many concgrams reveal patterns of collocation which would not have been uncovered, relying on intuition alone or other search engines. Concgram searches, by their very nature, emphasise the prevalence of word associations in language use, and diminish the attention that may be unduly paid to the node word(s) in user nominated queries in KWIC display. Such searches, we believe, will aid corpus linguists, and others in related fields, to uncover the full extent of the idiom principle (Sinclair, 1987).

References

Sinclair, J. McH. (1987). The nature of the evidence. In J. McH. Sinclair (ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins. 150-159.

[1] English Department, The Hong Kong Polytechnic University
e-mail: egwarren@polyu.edu.hk

Greaves, C. (2005). Introduction to ConcGram©. Tuscan Word Centre Workshop. Certosa di Pontignano, Tuscany, Italy, 25–29 June 2005.
Cheng, W., Greaves, C. and Warren, M. 2006. From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics* 11/4: 411–33.