Graph-based Analysis of Native and Learner Phraseology

Piotr PĘZIK
Institute of English Studies, PELCRA Group
University of Łódź
Łódź, Poland
piotr.pezik@uni.lodz.pl

One of the implications of recent research in phraseology is that "native-like proficiency in a language depends crucially on a stock of prefabricated units (…) varying in complexity and internal stability" (Cowie, 1998). This observation has been confirmed by large-scale studies conducted on reference corpora (partly reported in this paper), which have shown that, for example, most occurrences of noun phrases are lexically recurrent across different text genres and registers. In other words, most lexical realizations of naturally occurring noun phrases in English are potential lexical collocations, figurative and pure idioms, binomials and other types of reproduced word combinations (Paquot & Granger, 2008). This universal tendency for lexical items to form phraseological units and attain non-compositional meaning poses obvious problems for language learners. Although the development of phraseological competence has been the focus of several recent corpus-based SLA and language didactics studies, new approaches are needed to develop methods of identifying key problem areas in phraseology learning and phraseodidactics.

The present paper addresses this need by introducing a graph-based method of analysing learners' phraseological competence which relies on the automatic identification of phraseological word combinations found in learner and reference corpora. Potential lexical and grammatical collocations are first identified in the British National Corpus for a number of predefined syntactic patters using a positional-relational approach to collocation extraction. This step results in a reference online dictionary containing over 1.5 million lexical and grammatical collocations stored in a relational database. For each entry in the dictionary, detailed statistical profiles are computed, including information about the strength of association between the collocation constituents and their evenness of distribution. Entries from such dictionaries compiled from the BNC and the Polish English Learner Corpus (PLEC) are then used to generate graph-based visualisation of collocational profiles of lexical items. The visualisations are essentially undirected graphs in which each word is a vertex and each co-occurrence between distinct pairs of words is an undirected edge whose width depends on the statistical strength of a given co-occurrence or its evenness of distribution e.g. (Savický & Hlavácová, 2002). A collocational graph generated for a set of words found in the learner corpus is compared with a corresponding graph found in the reference corpus, thus revealing the differences and similarities between native and learner phraseological profiles. We have found this method to produce fairly accurate models of both learner and native phraseological competence for individual lexical items.

Apart from their applications in theoretical SLA research, collocational graphs created from a reference corpus such as the BNC can also be used as vocabulary teaching and exploration aids illustrating the use of apparently synonymous words which learners tend to misuse.

References

Cowie, A.P. (ed). (1998). *Phraseology Theory, Analysis, and Applications*. Oxford: Oxford University Press.

Paquot M. & Granger S. (2008). ‘Disentangling the phraseological web.’ In: Granger, S. & Meunier, F. (eds). *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins. 27-49.

Savický, P. & Hlavácová, J. (2002). ‘Measures of word commonness.’ *Journal of Quantitative Linguistics* 9:215-31.