Abs-189

Adriano Ferraresi (University of Naples “Federico II” / University of Bologna, Italy) and Stefan Th. Gries (University of California, Santa Barbara)

Type and (?) token frequencies in measures of collocational strength: lexical gravity vs. a few classics

Most well-established measures for calculating the collocational strength between words X and Y (e.g. MI, t-score and log-likelihood) are based on their joint frequency n and their two overall token frequencies a and b in a corpus. As such, they do not take into account how many different *types* co-occur with x in the position of y (and vice versa). However, this has been suggested to be a relevant criterion, e.g. within so-called "phraseological" approaches (Nesselhauf 2004) where "restricted collocations" are usually defined in terms of the number of different words (i.e. types) a node co-occurs with. Also, recent psycholinguistic studies show that e.g. the acquisition of syntactic patterns and their diachronic change are influenced by the diversity of their linguistic contexts, one operationalization of which are of course type frequencies (cf. Goldberg 2006, Bybee 2010).

Recently, Daudaravičius and Marcinkevičienė (2004) introduced lexical gravity, which incorporates information on frequency of type co-occurrence into its computation. Despite its potential theoretical interest, this measure is still underexplored in the corpus linguistics literature. Exceptions are Gries (2010), who finds that it outperforms t-score when used as a feature in cluster analysis to discriminate between different (sub-)registers in the BNC Baby, and Gries and Mukherjee (2010), who calculate lexical gravities for n-grams in different components of the ICE corpus revealing differences between varieties of Asian English and British English. However, lexical gravity has never been analysed per se as a measure of collocational strength and compared to better-established ones.

Relying on part-of-speech patterns for the identification of collocation candidates (along the lines of e.g. Evert 2008), this paper takes a first step towards filling this gap. Lexical gravity, MI, log-likelihood and bare frequency are calculated for adjective-noun pairs in two corpora: (i) a relatively small (~ 7m words) specialised corpus of English consisting of webpages of British and Irish universities (Bernardini et al. 2009), and (ii) the British National Corpus. The results are compared by means of rank correlations to explore the degrees to which the new measure returns different results than the others. In a second step we use a “stratified sampling” method, whereby for every measure we split the scored bigram lists into frequency ranges and extract the bigrams with the highest and lowest scores within these ranges, i.e. the bigrams for which, frequency being comparable, the most conflicting results are obtained. This method allows a more thorough scrutiny of the lists, since it does not limit itself to considering the n top-scored bigrams. Results suggest that lexical gravity is strongly correlated with co-occurrence frequency, but at the same time it distinguishes, within one frequency bin, between groups of salient and less salient bigrams.

References

Bernardini, S., A. Ferraresi and F. Gaspari. 2009. "Institutional English in Italian University websites: the acWaC corpus". Paper presented at *Corpus Linguistics 2009*, University of Liverpool.
Bybee, J. 2010. *Language, usage, and cognition*. Cambridge: Cambridge University Press.
Daudaravičius, V. and R. Marcinkevičienė. 2004. "Gravity counts for the boundaries of collocations". *International Journal of Corpus Linguistics* 9(2). 321-348.
Evert, S. 2008. "A lexicographic evaluation of German adjective-noun collocations". In *Proceedings of the LREC Workshop: Towards a Shared Task for Multiword Expressions*. Marrakech, Morocco.
Goldberg, A. E. 2006. *Constructions at work: on the nature of generalization in language*. Oxford: Oxford University Press.
Gries, S. Th. 2010. Bigrams in registers, domains, and varieties: a bigram gravity approach to the homogeneity of corpora. In *Proceedings of Corpus Linguistics 2009*, University of Liverpool.
Gries, S. Th. and J. Mukherjee. 2010. Lexical gravity across varieties of English: an ICE-based study of

n-grams in Asian Englishes. *International Journal of Corpus Linguistics* 15(4). 520-548.
Nesselhauf, N. 2004. *Collocations in a learner corpus*. Amsterdam: Benjamins.