Abs-235

Matthew Brook O'Donnell (University of Michigan)

The adjusted frequency list: A new method to extract cluster-sensitive frequency lists from corpora

The importance of multi-word chunks is well established in psycholinguistic (Erman & Warren 2000; Ellis 2003) and corpus linguistic (Sinclair 1991; O'Keeffe et al 2006) circles. However, many of our computational tools and methods still focus on individual words as the foundational units of analysis.

The adjusted frequency list is 'cluster sensitive' method for collecting n-grams, boosting the rank of larger word sequences ('on the other hand') and reduces the counts of their component parts ('on the', 'the other hand', etc.). Using sections of the BNCBaby corpus, a number of comparisons of unadjusted (standard) and adjusted frequency 1- to 5-gram lists are compared. For example, the top 10 items in a standard combined 1- to 5-gram frequency lists from BNCBaby-Demographic are: i, you, the, it, and, a, to, that, yeah, oh. The adjusted frequency method produces a cluster-sensitive list: i don't know, and, the, do you want, one two three, i don't think, of, in, two three four, a.

The simple method presented here, along with other more complex techniques that have been recently proposed (Gries & Mukherjee, 2010), demonstrates how corpus analysis continues to validate the importance of chunking in the investigation and description of language.

References

Ellis, N. C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. In C. Doughty & Long, M.H. (Eds.), Handbook of second language acquisition. Oxford: Blackwell: 33-68.

Erman, B. & Warren, B. (2000). The idiom principle and the open choice principle. Text 20 (1): 29-62.

Gries, S. & Mukherjee, J. (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.

O'Keeffe, A, McCarthy, M. & Carter, R. (2006). From Corpus to Classroom: Language Use and Language Teaching. Cambridge: Cambridge University Press.

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.