Extending the frequency measures: Bridging psycholinguistics to cognitive linguistics via corpus linguistics

Hien Pham
University of Alberta

Over the last 50 years, cognitive linguistics and psycholinguistics have relied on frequency of occurrence and demonstrated that language processing is sensitive to usage frequency at all levels of language representation (Ellis, 2002). This factor, one of the most robust predictors, has recently been challenged by the findings that frequency of occurrence explains only a modest proportion of lexical variability, while most of the variance in lexical spaces is explained by contextual measures (Baayen, 2010). These findings are also backed up by associative learning in cognitive psychology (Ellis, 2002). Constructions, form – meaning /function mappings, are regarded as language units. Acquisition of these units requires inducing associations from language experience. Constructionist approaches to language acquisition thus entail the distributional analysis of language data using statistical learning mechanism. The question addressed here is given that frequency is highly correlated with many contextual measures (e.g., dispersion, family size, entropy, among others presented below), what predictors aid the processing of compounds and word sequences?

The present research investigates various psycholinguistic factors which conspire in the acquisition and production of compound words and classifier tri-grams in Vietnamese. These factors include frequency, dispersion (also known as contextual diversity) (Adelman *et al.* 2006), Gries's Deviation of Proportion (DP) (Gries, 2009), Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997), Hyper Analog to Language (HAL) (Burgess & Lund, 1997). The above measures are all frequency-based measures, with the first three measures using lexical frequency (statistical measures) and the last two using co-occurrence frequency of words in the corpora (statistical-semantic measures).

Two corpora, a general corpus of newspaper articles and short stories, and a corpus of film subtitles, were constructed, tokenized, and tagged. These corpora were used to compute the statistical and statistical-semantic measures. Two visual lexical decision experiments were run on native Vietnamese readers: (1) a single-subject experiment with some 15000 compounds and the same number of nonwords, and (2) a multi-subject experiment (28 subjects) with 550 compound words and the same number of nonwords. We present the preliminary results showing the absolute correlations of the visual lexical decision data with the frequencies, dispersion measures, DP, LSA, and HAL.

We conclude with the discussion about the implications of frequency measures in cognitive linguistics and psycholinguistics. Cognitive linguistics uses type and token frequencies to correlates with morphological productivity and entrenchment respectively. Meanwhile psycholinguistics use type and token frequencies to predict the facilitatory or inhibitory effects in language comprehension/production. We suggest that using the statistical measures and the statistical-semantic measures presented above can bring them to an interdisciplinary development where many facets of the language complexity can be considered dynamically as the nature of language itself.

References

Adelman, J., Brown, G., & Quesada, J. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17*(9), 814.

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon, 5*(3), 436–461.

Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, *12*(2-3), 177–210.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–88.

Gries, S. (2009). Dispersions and adjusted frequencies in corpora: further explorations. *Language and Computers*, *71*(1), 197–212.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review, 104*(2), 211–240.